INTRODUCTORY ECONOMETRICS

A Modern Approach 7e

Jeffrey M. Wooldridge

Introductory Econometrics

A MODERN APPROACH

SEVENTH EDITION

Jeffrey M. Wooldridge
Michigan State University

CENGAGE

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

***Introductory Econometrics: A Modern Approach,* Seventh Edition**
Jeffrey M. Wooldridge

Senior Vice President, Higher Education Product Management: Erin Joyner

Product Director: Jason Fremder

Sr. Product Manager: Michael Parthenakis

Sr. Learning Designer: Sarah Keeling

Sr. Content Manager: Anita Verma

In-House Subject Matter Expert (s): Eugenia Belova, Ethan Crist and Kasie Jean

Digital Delivery Lead: Timothy Christy

Product Assistant: Matt Schiesl

Manufacturing Planner: Kevin Kluck

Production Service: SPi-Global

Intellectual Property

Analyst: Jennifer Bowes

Project Manager: Julie Geagan-Chevez

Marketing Manager: John Carey

Sr. Designer: Bethany Bourgeois

Cover Designer: Tin Box Studio

Library of Congress Control Number: 2018956380

ISBN: 978-1-337-55886-0

Cengage
20 Channel Center Street
Boston, MA 02210
USA

Cengage is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at **www.cengage.com**.

Cengage products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit **www.cengage.com**.

Printed in the United States of America
Print Number: 01 Print Year: 2018

Brief Contents

Contents

Preface

In ALL content, please indent the first paragraph as well, like the following ones. My motivation for writing the first edition of *Introductory Econometrics: A Modern Approach* was that I saw a fairly wide gap between how econometrics is taught to undergraduates and how empirical researchers think about and apply econometric methods. I became convinced that teaching introductory econometrics from the perspective of professional users of econometrics would actually simplify the presentation, in addition to making the subject much more interesting.

Based on the positive reactions to the several earlier editions, it appears that my hunch was correct. Many instructors, having a variety of backgrounds and interests and teaching students with different levels of preparation, have embraced the modern approach to econometrics espoused in this text. The emphasis in this edition is still on applying econometrics to real-world problems. Each econometric method is motivated by a particular issue facing researchers analyzing nonexperimental data. The focus in the main text is on understanding and interpreting the assumptions in light of actual empirical applications: the mathematics required is no more than college algebra and basic probability and statistics.

Designed for Today's Econometrics Course

The seventh edition preserves the overall organization of the sixth. The most noticeable feature that distinguishes this text from most others is the separation of topics by the kind of data being analyzed. This is a clear departure from the traditional approach, which presents a linear model, lists all assumptions that may be needed at some future point in the analysis, and then proves or asserts results without clearly connecting them to the assumptions. My approach is first to treat, in Part 1, multiple regression analysis with cross-sectional data, under the assumption of random sampling. This setting is natural to students because they are familiar with random sampling from a population in their introductory statistics courses. Importantly, it allows us to distinguish assumptions made about the underlying population regression model—assumptions that can be given economic or behavioral content—from assumptions about how the data were sampled. Discussions about the consequences of nonrandom sampling can be treated in an intuitive fashion after the students have a good grasp of the multiple regression model estimated using random samples.

An important feature of a modern approach is that the explanatory variables—along with the dependent variable—are treated as outcomes of random variables. For the social sciences, allowing random explanatory variables is much more realistic than the traditional assumption of nonrandom explanatory variables. As a nontrivial benefit, the population model/random sampling approach reduces the number of assumptions that students must absorb and understand. Ironically, the classical approach to regression analysis, which treats the explanatory variables as fixed in repeated samples and is still pervasive in introductory texts, literally applies to data collected in an experimental setting. In addition, the contortions required to state and explain assumptions can be confusing to students.

My focus on the population model emphasizes that the fundamental assumptions underlying regression analysis, such as the zero mean assumption on the unobservable error term, are properly

stated conditional on the explanatory variables. This leads to a clear understanding of the kinds of problems, such as heteroskedasticity (nonconstant variance), that can invalidate standard inference procedures. By focusing on the population, I am also able to dispel several misconceptions that arise in econometrics texts at all levels. For example, I explain why the usual R-squared is still valid as a goodness-of-fit measure in the presence of heteroskedasticity (Chapter 8) or serially correlated errors (Chapter 12); I provide a simple demonstration that tests for functional form should not be viewed as general tests of omitted variables (Chapter 9); and I explain why one should always include in a regression model extra control variables that are uncorrelated with the explanatory variable of interest, which is often a key policy variable (Chapter 6).

Because the assumptions for cross-sectional analysis are relatively straightforward yet realistic, students can get involved early with serious cross-sectional applications without having to worry about the thorny issues of trends, seasonality, serial correlation, high persistence, and spurious regression that are ubiquitous in time series regression models. Initially, I figured that my treatment of regression with cross-sectional data followed by regression with time series data would find favor with instructors whose own research interests are in applied microeconomics, and that appears to be the case. It has been gratifying that adopters of the text with an applied time series bent have been equally enthusiastic about the structure of the text. By postponing the econometric analysis of time series data, I am able to put proper focus on the potential pitfalls in analyzing time series data that do not arise with cross-sectional data. In effect, time series econometrics finally gets the serious treatment it deserves in an introductory text.

As in the earlier editions, I have consciously chosen topics that are important for reading journal articles and for conducting basic empirical research. Within each topic, I have deliberately omitted many tests and estimation procedures that, while traditionally included in textbooks, have not withstood the empirical test of time. Likewise, I have emphasized more recent topics that have clearly demonstrated their usefulness, such as obtaining test statistics that are robust to heteroskedasticity (or serial correlation) of unknown form, using multiple years of data for policy analysis, or solving the omitted variable problem by instrumental variables methods. I appear to have made fairly good choices, as I have received only a handful of suggestions for adding or deleting material.

I take a systematic approach throughout the text, by which I mean that each topic is presented by building on the previous material in a logical fashion, and assumptions are introduced only as they are needed to obtain a conclusion. For example, empirical researchers who use econometrics in their research understand that not all of the Gauss-Markov assumptions are needed to show that the ordinary least squares (OLS) estimators are unbiased. Yet the vast majority of econometrics texts introduce a complete set of assumptions (many of which are redundant or in some cases even logically conflicting) before proving the unbiasedness of OLS. Similarly, the normality assumption is often included among the assumptions that are needed for the Gauss-Markov Theorem, even though it is fairly well known that normality plays no role in showing that the OLS estimators are the best linear unbiased estimators.

My systematic approach is illustrated by the order of assumptions that I use for multiple regression in Part 1. This structure results in a natural progression for briefly summarizing the role of each assumption:

MLR.1: Introduce the population model and interpret the population parameters (which we hope to estimate).

MLR.2: Introduce random sampling from the population and describe the data that we use to estimate the population parameters.

MLR.3: Add the assumption on the explanatory variables that allows us to compute the estimates from our sample; this is the so-called no perfect collinearity assumption.

MLR.4: Assume that, in the population, the mean of the unobservable error does not depend on the values of the explanatory variables; this is the "mean independence" assumption combined with a zero population mean for the error, and it is the key assumption that delivers unbiasedness of OLS.

After introducing Assumptions MLR.1 to MLR.3, one can discuss the algebraic properties of ordinary least squares—that is, the properties of OLS for a particular set of data. By adding Assumption MLR.4, we can show that OLS is unbiased (and consistent). Assumption MLR.5 (homoskedasticity) is added for the Gauss-Markov Theorem and for the usual OLS variance formulas to be valid. Assumption MLR.6 (normality), which is not introduced until Chapter 4, is added to round out the classical linear model assumptions. The six assumptions are used to obtain exact statistical inference and to conclude that the OLS estimators have the smallest variances among all unbiased estimators.

I use parallel approaches when I turn to the study of large-sample properties and when I treat regression for time series data in Part 2. The careful presentation and discussion of assumptions makes it relatively easy to transition to Part 3, which covers advanced topics that include using pooled cross-sectional data, exploiting panel data structures, and applying instrumental variables methods. Generally, I have strived to provide a unified view of econometrics, where all estimators and test statistics are obtained using just a few intuitively reasonable principles of estimation and testing (which, of course, also have rigorous justification). For example, regression-based tests for heteroskedasticity and serial correlation are easy for students to grasp because they already have a solid understanding of regression. This is in contrast to treatments that give a set of disjointed recipes for outdated econometric testing procedures.

Throughout the text, I emphasize ceteris paribus relationships, which is why, after one chapter on the simple regression model, I move to multiple regression analysis. The multiple regression setting motivates students to think about serious applications early. I also give prominence to policy analysis with all kinds of data structures. Practical topics, such as using proxy variables to obtain ceteris paribus effects and interpreting partial effects in models with interaction terms, are covered in a simple fashion.

Designed at Undergraduates, Applicable to Master's Students

The text is designed for undergraduate economics majors who have taken college algebra and one-semester of introductory probability and statistics. (Math Refresher A, B, and C contain the requisite background material.) A one-semester or one-quarter econometrics course would not be expected to cover all, or even any, of the more advanced material in Part 3. A typical introductory course includes Chapters 1 through 8, which cover the basics of simple and multiple regression for cross-sectional data. Provided the emphasis is on intuition and interpreting the empirical examples, the material from the first eight chapters should be accessible to undergraduates in most economics departments. Most instructors will also want to cover at least parts of the chapters on regression analysis with time series data, Chapters 10 and 12, in varying degrees of depth. In the one-semester course that I teach at Michigan State, I cover Chapter 10 fairly carefully, give an overview of the material in Chapter 11, and cover the material on serial correlation in Chapter 12. I find that this basic one-semester course puts students on a solid footing to write empirical papers, such as a term paper, a senior seminar paper, or a senior thesis. Chapter 9 contains more specialized topics that arise in analyzing cross-sectional data, including data problems such as outliers and nonrandom sampling; for a one-semester course, it can be skipped without loss of continuity.

The structure of the text makes it ideal for a course with a cross-sectional or policy analysis focus: the time series chapters can be skipped in lieu of topics from Chapters 9 or 15. The new material on potential outcomes added to the first nine chapters should help the instructor craft a course that provides an introduction to modern policy analysis. Chapter 13 is advanced only in the sense that it treats two new data structures: independently pooled cross sections and two-period panel data analysis. Such data structures are especially useful for policy analysis, and the chapter provides

several examples. Students with a good grasp of Chapters 1 through 8 will have little difficulty with Chapter 13. Chapter 14 covers more advanced panel data methods and would probably be covered only in a second course. A good way to end a course on cross-sectional methods is to cover the rudiments of instrumental variables estimation in Chapter 15.

I have used selected material in Part 3, including Chapters 13 and 17, in a senior seminar geared to producing a serious research paper. Along with the basic one-semester course, students who have been exposed to basic panel data analysis, instrumental variables estimation, and limited dependent variable models are in a position to read large segments of the applied social sciences literature. Chapter 17 provides an introduction to the most common limited dependent variable models.

The text is also well suited for an introductory master's level course, where the emphasis is on applications rather than on derivations using matrix algebra. Several instructors have used the text to teach policy analysis at the master's level. For instructors wanting to present the material in matrix form, Appendices D and E are self-contained treatments of the matrix algebra and the multiple regression model in matrix form.

At Michigan State, PhD students in many fields that require data analysis—including accounting, agricultural economics, development economics, economics of education, finance, international economics, labor economics, macroeconomics, political science, and public finance—have found the text to be a useful bridge between the empirical work that they read and the more theoretical econometrics they learn at the PhD level.

Suggestions for Designing Your Course Beyond the Basic

I have already commented on the contents of most of the chapters as well as possible outlines for courses. Here I provide more specific comments about material in chapters that might be covered or skipped:

Chapter 9 has some interesting examples (such as a wage regression that includes IQ score as an explanatory variable). The rubric of proxy variables does not have to be formally introduced to present these kinds of examples, and I typically do so when finishing up cross-sectional analysis. In Chapter 12, for a one-semester course, I skip the material on serial correlation robust inference for ordinary least squares as well as dynamic models of heteroskedasticity.

Even in a second course I tend to spend only a little time on Chapter 16, which covers simultaneous equations analysis. I have found that instructors differ widely in their opinions on the importance of teaching simultaneous equations models to undergraduates. Some think this material is fundamental; others think it is rarely applicable. My own view is that simultaneous equations models are overused (see Chapter 16 for a discussion). If one reads applications carefully, omitted variables and measurement error are much more likely to be the reason one adopts instrumental variables estimation, and this is why I use omitted variables to motivate instrumental variables estimation in Chapter 15. Still, simultaneous equations models are indispensable for estimating demand and supply functions, and they apply in some other important cases as well.

Chapter 17 is the only chapter that considers models inherently nonlinear in their parameters, and this puts an extra burden on the student. The first material one should cover in this chapter is on probit and logit models for binary response. My presentation of Tobit models and censored regression still appears to be novel in introductory texts. I explicitly recognize that the Tobit model is applied to corner solution outcomes on random samples, while censored regression is applied when the data collection process censors the dependent variable at essentially arbitrary thresholds.

Chapter 18 covers some recent important topics from time series econometrics, including testing for unit roots and cointegration. I cover this material only in a second-semester course at either the undergraduate or master's level. A fairly detailed introduction to forecasting is also included in Chapter 18.

Chapter 19, which would be added to the syllabus for a course that requires a term paper, is much more extensive than similar chapters in other texts. It summarizes some of the methods appropriate for various kinds of problems and data structures, points out potential pitfalls, explains in some detail how to write a term paper in empirical economics, and includes suggestions for possible projects.

What's Changed?

I have added new exercises to many chapters, including to the Math Refresher and Advanced Treatment appendices. Some of the new computer exercises use new data sets, including a data set on performance of men's college basketball teams. I have also added more challenging problems that require derivations.

There are several notable changes to the text. An important organizational change, which should facilitate a wider variety of teaching tastes, is that the notion of binary, or dummy, explanatory variables is introduced in Chapter 2. There, it is shown that ordinary least squares estimation leads to a staple in basic statistics: the difference in means between two subgroups in a population. By introducing qualitative factors into regression early on, the instructor is able to use a wider variety of empirical examples from the very beginning.

The early discussion of binary explanatory variables allows for a formal introduction of potential, or counterfactual, outcomes, which is indispensable in the modern literature on estimating causal effects. The counterfactual approach to studying causality appears in previous editions, but Chapters 2, 3, 4, and 7 now explicitly include new sections on the modern approach to causal inference. Because basic policy analysis involves the binary decision to participate in a program or not, a leading example of using dummy independent variables in simple and multiple regression is to evaluate policy interventions. At the same time, the new material is incorporated into the text so that instructors not wishing to cover the potential outcomes framework may easily skip the material. Several end-of-chapter problems concern extensions of the basic potential outcomes framework, which should be valuable for instructors wishing to cover that material.

Chapter 3 includes a new section on different ways that one can apply multiple regression, including problems of pure prediction, testing efficient markets, and culminating with a discussion of estimating treatment or causal effects. I think this section provides a nice way to organize students' thinking about the scope of multiple regression after they have seen the mechanics of ordinary least squares (OS) and several examples. As with other new material that touches on causal effects, this material can be skipped without loss of continuity. A new section in Chapter 7 continues the discussion of potential outcomes, allowing for nonconstant treatment effects. The material is a nice illustration of estimating different regression functions for two subgroups from a population. New problems in this chapter that allow the student more experience in using full regression adjustment to estimate causal effects.

One notable change to Chapter 9 is a more detailed discussion of using missing data indicators when data are missing on one or more of the explanatory variables. The assumptions underlying the method are discussed in more detail than in the previous edition.

Chapter 12 has been reorganized to reflect a more modern treatment of the problem of serial correlation in the errors of time series regression models. The new structure first covers adjusting the OLS standard errors to allow general forms of serial correlation. Thus, the chapter outline now parallels that in Chapter 8, with the emphasis in both cases on OLS estimation but making inference robust to violation of standard assumptions. Correcting for serial correlation using generalized least squares now comes after OLS and the treatment of testing for serial correlation.

The advanced chapters also include several improvements. Chapter 13 now discusses, at an accessible level, extensions of the standard difference-in-differences setup, allowing for multiple control

groups, multiple time periods, and even group-specific trends. In addition, the chapter includes a more detailed discussion of computing standard errors robust to serial correlation when using first-differencing estimation with panel data.

Chapter 14 now provides more detailed discussions of several important issues in estimating panel data models by fixed effects, random effects, and correlated random effects (CRE). The CRE approach with missing data is discussed in more detail, as is how one accounts for general functional forms, such as squares and interactions, which are covered in the cross-sectional setting in Chapter 6. An expanded section on general policy analysis with panel data should be useful for courses with an emphasis on program interventions and policy evaluation.

Chapter 16, which still covers simultaneous equations models, now provides an explicit link between the potential outcomes framework and specification of simultaneous equations models.

Chapter 17 now includes a discussion of using regression adjustment for estimating causal (treatment) effects when the outcome variable has special features, such as when the outcome itself is a binary variable. Then, as the reader is asked to explore in a new problem, logit and probit models can be used to obtain more reliable estimates of average treatment effects by estimating separate models for each treatment group.

Chapter 18 now provides more details about how one can compute a proper standard error for a forecast (as opposed to a prediction) interval. This should help the advanced reader understand in more detail the nature of the uncertainty in the forecast.

About MindTap™

MindTap is an outcome-driven application that propels students from memorization to mastery. It's the only platform that gives you complete ownership of your course. With it, you can challenge every student, build their confidence, and empower them to be unstoppable.

Access Everything You Need In One Place. Cut down on prep with preloaded, organized course materials in MindTap. Teach more efficiently with interactive multimedia, assignments, quizzes and more. And give your students the power to read, listen and study on their phones, so they can learn on their terms.

Empower Your Students To Reach Their Potential. Twelve distinct metrics give you actionable insights into student engagement. Identify topics troubling your entire class and instantly communicate with struggling students. And students can track their scores to stay motivated toward their goals. Together, you can accelerate progress.

Your Course. Your Content. Only MindTap gives you complete control over your course. You have the flexibility to reorder textbook chapters, add your own notes and embed a variety of content including OER. Personalize course content to your students' needs. They can even read your notes, add their own and highlight key text to aid their progress.

A Dedicated Team, Whenever You Need Them. MindTap isn't just a tool; it's backed by a personalized team eager to support you. Get help setting up your course and tailoring it to your specific objectives. You'll be ready to make an impact from day one. And, we'll be right here to help you and your students throughout the semester—and beyond.

Design Features

In addition to the didactic material in the chapter, I have included two features to help students better understand and apply what they are learning. Each chapter contains many numbered examples. Several of these are case studies drawn from recently published papers. I have used my judgment to simplify the analysis, hopefully without sacrificing the main point. The "Going Further Questions" in

the chapter provide students an opportunity to "go further" in learning the material through analysis or application. Students will find immediate feedback for these questions in the end of the text.

The end-of-chapter problems and computer exercises are heavily oriented toward empirical work, rather than complicated derivations. The students are asked to reason carefully based on what they have learned. The computer exercises often expand on the in-text examples. Several exercises use data sets from published works or similar data sets that are motivated by published research in economics and other fields.

A pioneering feature of this introductory econometrics text is the extensive glossary. The short definitions and descriptions are a helpful refresher for students studying for exams or reading empirical research that uses econometric methods. I have added and updated several entries for the seventh edition.

Instructional Tools

Cengage offers various supplements for instructors and students who use this book. I would like to thank the Subject Matter Expert team who worked on these supplements and made teaching and learning easy.

C. Patrick Scott, Ph.D., Louisiana Tech University (R Videos and Computer exercise reviewer)
Hisham Foad (Aplia Home work reviewer and Glossary)
Kenneth H. Brown, Missouri State University (R Videos creator)
Scott Kostyshak, University of Florida (R Videos reviewer)
Ujwal Kharel (Test Bank and Adaptive Test Prep)

Data Sets—Available in Six Formats

With more than 100 data sets in six different formats, including Stata®, R, EViews®, Minitab®, Microsoft® Excel, and Text, the instructor has many options for problem sets, examples, and term projects. Because most of the data sets come from actual research, some are very large. Except for partial lists of data sets to illustrate the various data structures, the data sets are not reported in the text. This book is geared to a course where computer work plays an integral role.

Updated Data Sets Handbook

An extensive data description manual is also available online. This manual contains a list of data sources along with suggestions for ways to use the data sets that are not described in the text. This unique handbook, created by author Jeffrey M. Wooldridge, lists the source of all data sets for quick reference and how each might be used. Because the data book contains page numbers, it is easy to see how the author used the data in the text. Students may want to view the descriptions of each data set and it can help guide instructors in generating new homework exercises, exam problems, or term projects. The author also provides suggestions on improving the data sets in this detailed resource that is available on the book's companion website at http://login.cengage.com and students can access it free at www.cengage.com.

Instructor's Manual with Solutions

REVISED INSTRUCTOR'S MANUAL WITH SOLUTIONS SAVES TIME IN PREPARATION AND GRADING. The online Instructor's Manual with solutions contains answers to all exercises in this edition. Teaching tips provide suggestions for presenting each chapter's material. The Instructor's Manual also contains sources for each of the data files with suggestions for using the data to develop problem sets, exams, and term papers. The Instructor's Manual is password-protected and available for download on the book's companion website.

Test Bank

Cengage Testing, powered by Cognero® is a flexible, online system that allows you to import, edit, and manipulate content from the text's test bank or elsewhere, including your own favorite test questions; create multiple test versions in an instant; and deliver tests from your LMS, your classroom, or wherever you want.

PowerPoint Slides

UPDATED POWERPOINT® SLIDES BRING LECTURES TO LIFE WHILE VISUALLY CLARIFYING CONCEPTS. Exceptional PowerPoint® presentation slides, created specifically for this edition, help you create engaging, memorable lectures. The slides are particularly useful for clarifying advanced topics in Part 3. You can modify or customize the slides for your specific course. PowerPoint® slides are available for convenient download on the instructor-only, password-protected section of the book's companion website.

Scientific Word Slides

UPDATED SCIENTIFIC WORD® SLIDES REINFORCE TEXT CONCEPTS AND LECTURE PRESENTATIONS. Created by the text author, this edition's Scientific Word® slides reinforce the book's presentation slides while highlighting the benefits of Scientific Word®, the application created by MacKichan software, Inc. for specifically composing mathematical, scientific and technical documents using LaTeX typesetting. These slides are based on the author's actual lectures and are available for convenient download on the password-protected section of the book's companion website.

Student Supplements

Student Solutions Manual

Now your student's can maximize their study time and further their course success with this dynamic online resource. This helpful Solutions Manual includes detailed steps and solutions to odd-numbered problems as well as computer exercises in the text. This supplement is available as a free resource at www.cengagebrain.com.

Acknowledgments

I would like to thank those who reviewed and provided helpful comments for this and previous editions of the text:

Erica Johnson, *Gonzaga University*

Mary Ellen Benedict, *Bowling Green State University*

Chirok Han, *Korea University*

Yan Li, *Temple University*

Melissa Tartari, *Yale University*

Michael Allgrunn, *University of South Dakota*

Gregory Colman, *Pace University*

Yoo-Mi Chin, *Missouri University of Science and Technology*

Arsen Melkumian, *Western Illinois University*

Kevin J. Murphy, *Oakland University*

Kristine Grimsrud, *University of New Mexico*

Will Melick, *Kenyon College*

Philip H. Brown, *Colby College*

Argun Saatcioglu, *University of Kansas*

Ken Brown, *University of Northern Iowa*

Michael R. Jonas, *University of San Francisco*

Melissa Yeoh, *Berry College*

Nikolaos Papanikolaou, *SUNY at New Paltz*

Konstantin Golyaev, *University of Minnesota*

Soren Hauge, *Ripon College*

Kevin Williams, *University of Minnesota*

Hailong Qian, *Saint Louis University*

Rod Hissong, *University of Texas at Arlington*

Steven Cuellar, *Sonoma State University*

Yanan Di, *Wagner College*

John Fitzgerald, *Bowdoin College*

Philip N. Jefferson, *Swarthmore College*

Yongsheng Wang, *Washington and Jefferson College*

Sheng-Kai Chang, *National Taiwan University*

Damayanti Ghosh, *Binghamton University*

Susan Averett, *Lafayette College*

Kevin J. Mumford, *Purdue University*

Nicolai V. Kuminoff, *Arizona State University*

Subarna K. Samanta, *The College of New Jersey*

Jing Li, *South Dakota State University*

Gary Wagner, *University of Arkansas–Little Rock*

Kelly Cobourn, *Boise State University*

Timothy Dittmer, *Central Washington University*

Daniel Fischmar, *Westminster College*

Subha Mani, *Fordham University*

John Maluccio, *Middlebury College*

James Warner, *College of Wooster*

Christopher Magee, *Bucknell University*

Andrew Ewing, *Eckerd College*

Debra Israel, *Indiana State University*

Jay Goodliffe, *Brigham Young University*

Stanley R. Thompson, *The Ohio State University*

Michael Robinson, *Mount Holyoke College*

Ivan Jeliazkov, *University of California, Irvine*

Heather O'Neill, *Ursinus College*

Leslie Papke, *Michigan State University*

Timothy Vogelsang, *Michigan State University*

Stephen Woodbury, *Michigan State University*

Some of the changes I discussed earlier were driven by comments I received from people on this list, and I continue to mull over other specific suggestions made by one or more reviewers.

Many students and teaching assistants, too numerous to list, have caught mistakes in earlier editions or have suggested rewording some paragraphs. I am grateful to them.

As always, it was a pleasure working with the team at Cengage Learning. Michael Parthenakis, my longtime Product Manager, has learned very well how to guide me with a firm yet gentle hand. Anita Verma and Ethan Crist quickly mastered the difficult challenges of being the content and subject matter expert team of a dense, technical textbook. Their careful reading of the manuscript and fine eye for detail have improved this seventh edition considerably.

This book is dedicated to my family: Leslie, Edmund, and R.G.

Jeffrey M. Wooldridge

About the Author

Jeffrey M. Wooldridge is University Distinguished Professor of Economics at Michigan State University, where he has taught since 1991. From 1986 to 1991, he was an assistant professor of economics at the Massachusetts Institute of Technology. He received his bachelor of arts, with majors in computer science and economics, from the University of California, Berkeley, in 1982, and received his doctorate in economics in 1986 from the University of California, San Diego. He has published more than 60 articles in internationally recognized journals, as well as several book chapters. He is also the author of *Econometric Analysis of Cross Section and Panel Data*, second edition. His awards include an Alfred P. Sloan Research Fellowship, the Plura Scripsit award from *Econometric Theory*, the Sir Richard Stone prize from the *Journal of Applied Econometrics*, and three graduate teacher-of-the-year awards from MIT. He is a fellow of the Econometric Society and of the *Journal of Econometrics*. He is past editor of the *Journal of Business and Economic Statistics*, and past econometrics coeditor of *Economics Letters*. He has served on the editorial boards of *Econometric Theory*, the *Journal of Economic Literature*, the *Journal of Econometrics*, the *Review of Economics and Statistics*, and the *Stata Journal*. He has also acted as an occasional econometrics consultant for Arthur Andersen, Charles River Associates, the Washington State Institute for Public Policy, Stratus Consulting, and Industrial Economics, Incorporated.

CHAPTER 1

The Nature of Econometrics and Economic Data

Chapter 1 discusses the scope of econometrics and raises general issues that arise in the application of econometric methods. Section 1-1 provides a brief discussion about the purpose and scope of econometrics and how it fits into economic analysis. Section 1-2 provides examples of how one can start with an economic theory and build a model that can be estimated using data. Section 1-3 examines the kinds of data sets that are used in business, economics, and other social sciences. Section 1-4 provides an intuitive discussion of the difficulties associated with inferring causality in the social sciences.

1-1 What Is Econometrics?

Imagine that you are hired by your state government to evaluate the effectiveness of a publicly funded job training program. Suppose this program teaches workers various ways to use computers in the manufacturing process. The 20-week program offers courses during nonworking hours. Any hourly manufacturing worker may participate, and enrollment in all or part of the program is voluntary. You are to determine what, if any, effect the training program has on each worker's subsequent hourly wage.

Now, suppose you work for an investment bank. You are to study the returns on different investment strategies involving short-term U.S. treasury bills to decide whether they comply with implied economic theories.

The task of answering such questions may seem daunting at first. At this point, you may only have a vague idea of the kind of data you would need to collect. By the end of this introductory econometrics course, you should know how to use econometric methods to formally evaluate a job training program or to test a simple economic theory.

Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy. A common application of econometrics is the forecasting of such important macroeconomic variables as interest rates, inflation rates, and gross domestic product (GDP). Whereas forecasts of economic indicators are highly visible and often widely published, econometric methods can be used in economic areas that have nothing to do with macroeconomic forecasting. For example, we will study the effects of political campaign expenditures on voting outcomes. We will consider the effect of school spending on student performance in the field of education. In addition, we will learn how to use econometric methods for forecasting economic time series.

Econometrics has evolved as a separate discipline from mathematical statistics because the former focuses on the problems inherent in collecting and analyzing nonexperimental economic data. **Nonexperimental data** are not accumulated through controlled experiments on individuals, firms, or segments of the economy. (Nonexperimental data are sometimes called **observational data**, or **retrospective data**, to emphasize the fact that the researcher is a passive collector of the data.) **Experimental data** are often collected in laboratory environments in the natural sciences, but they are more difficult to obtain in the social sciences. Although some social experiments can be devised, it is often impossible, prohibitively expensive, or morally repugnant to conduct the kinds of controlled experiments that would be needed to address economic issues. We give some specific examples of the differences between experimental and nonexperimental data in Section 1-4.

Naturally, econometricians have borrowed from mathematical statisticians whenever possible. The method of multiple regression analysis is the mainstay in both fields, but its focus and interpretation can differ markedly. In addition, economists have devised new techniques to deal with the complexities of economic data and to test the predictions of economic theories.

1-2 Steps in Empirical Economic Analysis

Econometric methods are relevant in virtually every branch of applied economics. They come into play either when we have an economic theory to test or when we have a relationship in mind that has some importance for business decisions or policy analysis. An **empirical analysis** uses data to test a theory or to estimate a relationship.

How does one go about structuring an empirical economic analysis? It may seem obvious, but it is worth emphasizing that the first step in any empirical analysis is the careful formulation of the question of interest. The question might deal with testing a certain aspect of an economic theory, or it might pertain to testing the effects of a government policy. In principle, econometric methods can be used to answer a wide range of questions.

In some cases, especially those that involve the testing of economic theories, a formal **economic model** is constructed. An economic model consists of mathematical equations that describe various relationships. Economists are well known for their building of models to describe a vast array of behaviors. For example, in intermediate microeconomics, individual consumption decisions, subject to a budget constraint, are described by mathematical models. The basic premise underlying these models is *utility maximization*. The assumption that individuals make choices to maximize their well-being, subject to resource constraints, gives us a very powerful framework for creating tractable economic models and making clear predictions. In the context of consumption decisions, utility maximization leads to a set of *demand equations*. In a demand equation, the quantity demanded of each commodity depends on the price of the goods, the price of substitute and complementary goods, the consumer's income, and the individual's characteristics that affect taste. These equations can form the basis of an econometric analysis of consumer demand.

Economists have used basic economic tools, such as the utility maximization framework, to explain behaviors that at first glance may appear to be noneconomic in nature. A classic example is Becker's (1968) economic model of criminal behavior.

EXAMPLE 1.1 **Economic Model of Crime**

In a seminal article, Nobel Prize winner Gary Becker postulated a utility maximization framework to describe an individual's participation in crime. Certain crimes have clear economic rewards, but most criminal behaviors have costs. The opportunity costs of crime prevent the criminal from participating in other activities such as legal employment. In addition, there are costs associated with the possibility of being caught and then, if convicted, the costs associated with incarceration. From Becker's perspective, the decision to undertake illegal activity is one of resource allocation, with the benefits and costs of competing activities taken into account.

Under general assumptions, we can derive an equation describing the amount of time spent in criminal activity as a function of various factors. We might represent such a function as

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7), \qquad [1.1]$$

where

y = hours spent in criminal activities,
x_1 = "wage" for an hour spent in criminal activity,
x_2 = hourly wage in legal employment,
x_3 = income other than from crime or employment,
x_4 = probability of getting caught,
x_5 = probability of being convicted if caught,
x_6 = expected sentence if convicted, and
x_7 = age.

Other factors generally affect a person's decision to participate in crime, but the list above is representative of what might result from a formal economic analysis. As is common in economic theory, we have not been specific about the function $f(\cdot)$ in (1.1). This function depends on an underlying utility function, which is rarely known. Nevertheless, we can use economic theory—or introspection—to predict the effect that each variable would have on criminal activity. This is the basis for an econometric analysis of individual criminal activity.

Formal economic modeling is sometimes the starting point for empirical analysis, but it is more common to use economic theory less formally, or even to rely entirely on intuition. You may agree that the determinants of criminal behavior appearing in equation (1.1) are reasonable based on common sense; we might arrive at such an equation directly, without starting from utility maximization. This view has some merit, although there are cases in which formal derivations provide insights that intuition can overlook.

Next is an example of an equation that we can derive through somewhat informal reasoning.

EXAMPLE 1.2 **Job Training and Worker Productivity**

Consider the problem posed at the beginning of Section 1-1. A labor economist would like to examine the effects of job training on worker productivity. In this case, there is little need for formal economic theory. Basic economic understanding is sufficient for realizing that factors such as education, experience, and training affect worker productivity. Also, economists are well aware that workers are paid commensurate with their productivity. This simple reasoning leads to a model such as

$$wage = f(educ, exper, training), \qquad [1.2]$$

where

wage = hourly wage,
educ = years of formal education,
exper = years of workforce experience, and
training = weeks spent in job training.

Again, other factors generally affect the wage rate, but equation (1.2) captures the essence of the problem.

After we specify an economic model, we need to turn it into what we call an **econometric model**. Because we will deal with econometric models throughout this text, it is important to know how an econometric model relates to an economic model. Take equation (1.1) as an example. The form of the function $f(\cdot)$ must be specified before we can undertake an econometric analysis. A second issue concerning (1.1) is how to deal with variables that cannot reasonably be observed. For example, consider the wage that a person can earn in criminal activity. In principle, such a quantity is well defined, but it would be difficult if not impossible to observe this wage for a given individual. Even variables such as the probability of being arrested cannot realistically be obtained for a given individual, but at least we can observe relevant arrest statistics and derive a variable that approximates the probability of arrest. Many other factors affect criminal behavior that we cannot even list, let alone observe, but we must somehow account for them.

The ambiguities inherent in the economic model of crime are resolved by specifying a particular econometric model:

$$crime = \beta_0 + \beta_1 wage + \beta_2 othinc + \beta_3 freqarr + \beta_4 freqconv + \beta_5 avgsen + \beta_6 age + u, \quad [1.3]$$

where

$crime$ = some measure of the frequency of criminal activity,
$wage$ = the wage that can be earned in legal employment,
$othinc$ = the income from other sources (assets, inheritance, and so on),
$freqarr$ = the frequency of arrests for prior infractions (to approximate the probability of arrest),
$freqconv$ = the frequency of conviction, and
$avgsen$ = the average sentence length after conviction.

The choice of these variables is determined by the economic theory as well as data considerations. The term u contains unobserved factors, such as the wage for criminal activity, moral character, family background, and errors in measuring things like criminal activity and the probability of arrest. We could add family background variables to the model, such as number of siblings, parents' education, and so on, but we can never eliminate u entirely. In fact, dealing with this *error term* or *disturbance term* is perhaps the most important component of any econometric analysis.

The constants $\beta_0, \beta_1, \ldots, \beta_6$ are the *parameters* of the econometric model, and they describe the directions and strengths of the relationship between *crime* and the factors used to determine *crime* in the model.

A complete econometric model for Example 1.2 might be

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u, \quad [1.4]$$

where the term u contains factors such as "innate ability," quality of education, family background, and the myriad other factors that can influence a person's wage. If we are specifically concerned about the effects of job training, then β_3 is the parameter of interest.

For the most part, econometric analysis begins by specifying an econometric model, without consideration of the details of the model's creation. We generally follow this approach, largely because careful derivation of something like the economic model of crime is time consuming and can take us into some specialized and often difficult areas of economic theory. Economic reasoning will play a role in our examples, and we will merge any underlying economic theory into the econometric model specification. In the economic model of crime example, we would start with an econometric model such as (1.3) and use economic reasoning and common sense as guides for choosing the variables. Although this approach loses some of the richness of economic analysis, it is commonly and effectively applied by careful researchers.

Once an econometric model such as (1.3) or (1.4) has been specified, various *hypotheses* of interest can be stated in terms of the unknown parameters. For example, in equation (1.3), we might hypothesize that *wage*, the wage that can be earned in legal employment, has no effect on criminal behavior. In the context of this particular econometric model, the hypothesis is equivalent to $\beta_1 = 0$.

An empirical analysis, by definition, requires data. After data on the relevant variables have been collected, econometric methods are used to estimate the parameters in the econometric model and to formally test hypotheses of interest. In some cases, the econometric model is used to make predictions in either the testing of a theory or the study of a policy's impact.

Because data collection is so important in empirical work, Section 1-3 will describe the kinds of data that we are likely to encounter.

1-3 The Structure of Economic Data

Economic data sets come in a variety of types. Whereas some econometric methods can be applied with little or no modification to many different kinds of data sets, the special features of some data sets must be accounted for or should be exploited. We next describe the most important data structures encountered in applied work.

1-3a Cross-Sectional Data

A **cross-sectional data set** consists of a sample of individuals, households, firms, cities, states, countries, or a variety of other units, taken at a given point in time. Sometimes, the data on all units do not correspond to precisely the same time period. For example, several families may be surveyed during different weeks within a year. In a pure cross-sectional analysis, we would ignore any minor timing differences in collecting the data. If a set of families was surveyed during different weeks of the same year, we would still view this as a cross-sectional data set.

An important feature of cross-sectional data is that we can often assume that they have been obtained by **random sampling** from the underlying population. For example, if we obtain information on wages, education, experience, and other characteristics by randomly drawing 500 people from the working population, then we have a random sample from the population of all working people. Random sampling is the sampling scheme covered in introductory statistics courses, and it simplifies the analysis of cross-sectional data. A review of random sampling is contained in Math Refresher C.

Sometimes, random sampling is not appropriate as an assumption for analyzing cross-sectional data. For example, suppose we are interested in studying factors that influence the accumulation of family wealth. We could survey a random sample of families, but some families might refuse to report their wealth. If, for example, wealthier families are less likely to disclose their wealth, then the resulting sample on wealth is not a random sample from the population of all families. This is an illustration of a sample selection problem, an advanced topic that we will discuss in Chapter 17.

Another violation of random sampling occurs when we sample from units that are large relative to the population, particularly geographical units. The potential problem in such cases is that the population is not large enough to reasonably assume the observations are independent draws. For example, if we want to explain new business activity across states as a function of wage rates, energy prices, corporate and property tax rates, services provided, quality of the workforce, and other state characteristics, it is unlikely that business activities in states near one another are independent. It turns out that the econometric methods that we discuss do work in such situations, but they sometimes need to be refined. For the most part, we will ignore the intricacies that arise in analyzing such situations and treat these problems in a random sampling framework, even when it is not technically correct to do so.

Cross-sectional data are widely used in economics and other social sciences. In economics, the analysis of cross-sectional data is closely aligned with the applied microeconomics fields, such as labor economics, state and local public finance, industrial organization, urban economics, demography, and health economics. Data on individuals, households, firms, and cities at a given point in time are important for testing microeconomic hypotheses and evaluating economic policies.

The cross-sectional data used for econometric analysis can be represented and stored in computers. Table 1.1 contains, in abbreviated form, a cross-sectional data set on 526 working individuals

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

for the year 1976. (This is a subset of the data in the file WAGE1.) The variables include *wage* (in dollars per hour), *educ* (years of education), *exper* (years of potential labor force experience), *female* (an indicator for gender), and *married* (marital status). These last two variables are binary (zero-one) in nature and serve to indicate qualitative features of the individual (the person is female or not; the person is married or not). We will have much to say about binary variables in Chapter 7 and beyond.

The variable *obsno* in Table 1.1 is the observation number assigned to each person in the sample. Unlike the other variables, it is not a characteristic of the individual. All econometrics and statistics software packages assign an observation number to each data unit. Intuition should tell you that, for data such as that in Table 1.1, it does not matter which person is labeled as observation 1, which person is called observation 2, and so on. The fact that the ordering of the data does not matter for econometric analysis is a key feature of cross-sectional data sets obtained from random sampling.

Different variables sometimes correspond to different time periods in cross-sectional data sets. For example, to determine the effects of government policies on long-term economic growth, economists have studied the relationship between growth in real per capita GDP over a certain period (say, 1960 to 1985) and variables determined in part by government policy in 1960 (government consumption as a percentage of GDP and adult secondary education rates). Such a data set might be represented as in Table 1.2, which constitutes part of the data set used in the study of cross-country growth rates by De Long and Summers (1991).

The variable *gpcrgdp* represents average growth in real per capita GDP over the period 1960 to 1985. The fact that *govcons60* (government consumption as a percentage of GDP) and *second60*

TABLE 1.2 A Data Set on Economic Growth Rates and Country Characteristics

obsno	country	gpcrgdp	govcons60	second60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Belgium	2.56	13	69
4	Bolivia	1.24	18	12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
61	Zimbabwe	2.30	17	6

(percentage of adult population with a secondary education) correspond to the year 1960, while *gpcrgdp* is the average growth over the period from 1960 to 1985, does not lead to any special problems in treating this information as a cross-sectional data set. The observations are listed alphabetically by country, but nothing about this ordering affects any subsequent analysis.

1-3b Time Series Data

A **time series data** set consists of observations on a variable or several variables over time. Examples of time series data include stock prices, money supply, consumer price index, GDP, annual homicide rates, and automobile sales figures. Because past events can influence future events and lags in behavior are prevalent in the social sciences, time is an important dimension in a time series data set. Unlike the arrangement of cross-sectional data, the chronological ordering of observations in a time series conveys potentially important information.

A key feature of time series data that makes them more difficult to analyze than cross-sectional data is that economic observations can rarely, if ever, be assumed to be independent across time. Most economic and other time series are related, often strongly related, to their recent histories. For example, knowing something about the GDP from last quarter tells us quite a bit about the likely range of the GDP during this quarter, because GDP tends to remain fairly stable from one quarter to the next. Although most econometric procedures can be used with both cross-sectional and time series data, more needs to be done in specifying econometric models for time series data before standard econometric methods can be justified. In addition, modifications and embellishments to standard econometric techniques have been developed to account for and exploit the dependent nature of economic time series and to address other issues, such as the fact that some economic variables tend to display clear trends over time.

Another feature of time series data that can require special attention is the **data frequency** at which the data are collected. In economics, the most common frequencies are daily, weekly, monthly, quarterly, and annually. Stock prices are recorded at daily intervals (excluding Saturday and Sunday). The money supply in the U.S. economy is reported weekly. Many macroeconomic series are tabulated monthly, including inflation and unemployment rates. Other macro series are recorded less frequently, such as every three months (every quarter). GDP is an important example of a quarterly series. Other time series, such as infant mortality rates for states in the United States, are available only on an annual basis.

Many weekly, monthly, and quarterly economic time series display a strong seasonal pattern, which can be an important factor in a time series analysis. For example, monthly data on housing starts differ across the months simply due to changing weather conditions. We will learn how to deal with seasonal time series in Chapter 10.

Table 1.3 contains a time series data set obtained from an article by Castillo-Freeman and Freeman (1992) on minimum wage effects in Puerto Rico. The earliest year in the data set is the first

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

observation, and the most recent year available is the last observation. When econometric methods are used to analyze time series data, the data should be stored in chronological order.

The variable *avgmin* refers to the average minimum wage for the year, *avgcov* is the average coverage rate (the percentage of workers covered by the minimum wage law), *prunemp* is the unemployment rate, and *prgnp* is the gross national product, in millions of 1954 dollars. We will use these data later in a time series analysis of the effect of the minimum wage on employment.

1-3c Pooled Cross Sections

Some data sets have both cross-sectional and time series features. For example, suppose that two cross-sectional household surveys are taken in the United States, one in 1985 and one in 1990. In 1985, a random sample of households is surveyed for variables such as income, savings, family size, and so on. In 1990, a *new* random sample of households is taken using the same survey questions. To increase our sample size, we can form a **pooled cross section** by combining the two years.

Pooling cross sections from different years is often an effective way of analyzing the effects of a new government policy. The idea is to collect data from the years before and after a key policy change. As an example, consider the following data set on housing prices taken in 1993 and 1995, before and after a reduction in property taxes in 1994. Suppose we have data on 250 houses for 1993 and on 270 houses for 1995. One way to store such a data set is given in Table 1.4.

Observations 1 through 250 correspond to the houses sold in 1993, and observations 251 through 520 correspond to the 270 houses sold in 1995. Although the order in which we store the data turns out not to be crucial, keeping track of the year for each observation is usually very important. This is why we enter *year* as a separate variable.

A pooled cross section is analyzed much like a standard cross section, except that we often need to account for secular differences in the variables across the time. In fact, in addition to increasing the sample size, the point of a pooled cross-sectional analysis is often to see how a key relationship has changed over time.

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85,500	42	1600	3	2.0
2	1993	67,300	36	1440	3	2.5
3	1993	134,000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243,600	41	2600	4	3.0
251	1995	65,000	16	1250	2	1.0
252	1995	182,400	20	2200	4	2.0
253	1995	97,500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57,200	16	1100	2	1.5

1-3d Panel or Longitudinal Data

A **panel data** (or *longitudinal data*) set consists of a time series for *each* cross-sectional member in the data set. As an example, suppose we have wage, education, and employment history for a set of individuals followed over a 10-year period. Or we might collect information, such as investment and financial data, about the same set of firms over a five-year time period. Panel data can also be collected on geographical units. For example, we can collect data for the same set of counties in the United States on immigration flows, tax rates, wage rates, government expenditures, and so on, for the years 1980, 1985, and 1990.

The key feature of panel data that distinguishes them from a pooled cross section is that the *same* cross-sectional units (individuals, firms, or counties in the preceding examples) are followed over a given time period. The data in Table 1.4 are not considered a panel data set because the houses sold are likely to be different in 1993 and 1995; if there are any duplicates, the number is likely to be so small as to be unimportant. In contrast, Table 1.5 contains a two-year panel data set on crime and related statistics for 150 cities in the United States.

There are several interesting features in Table 1.5. First, each city has been given a number from 1 through 150. Which city we decide to call city 1, city 2, and so on, is irrelevant. As with a pure cross section, the ordering in the cross section of a panel data set does not matter. We could use the city name in place of a number, but it is often useful to have both.

A second point is that the two years of data for city 1 fill the first two rows or observations, observations 3 and 4 correspond to city 2, and so on. Because each of the 150 cities has two rows of data, any econometrics package will view this as 300 observations. This data set can be treated as a pooled cross section, where the same cities happen to show up in each year. But, as we will see in Chapters 13 and 14, we can also use the panel structure to analyze questions that cannot be answered by simply viewing this as a pooled cross section.

In organizing the observations in Table 1.5, we place the two years of data for each city adjacent to one another, with the first year coming before the second in all cases. For just about every practical purpose, this is the preferred way for ordering panel data sets. Contrast this organization with the way the pooled cross sections are stored in Table 1.4. In short, the reason for ordering panel data as in Table 1.5 is that we will need to perform data transformations for each city across the two years.

Because panel data require replication of the same units over time, panel data sets, especially those on individuals, households, and firms, are more difficult to obtain than pooled cross sections. Not surprisingly, observing the same units over time leads to several advantages over cross-sectional data or even pooled cross-sectional data. The benefit that we will focus on in this text is that having

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350,000	8.7	440
2	1	1990	8	359,200	7.2	471
3	2	1986	2	64,300	5.4	75
4	2	1990	1	65,100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260,700	9.6	286
298	149	1990	6	245,000	9.8	334
299	150	1986	25	543,000	4.3	520
300	150	1990	32	546,200	5.2	493

multiple observations on the same units allows us to control for certain unobserved characteristics of individuals, firms, and so on. As we will see, the use of more than one observation can facilitate causal inference in situations where inferring causality would be very difficult if only a single cross section were available. A second advantage of panel data is that they often allow us to study the importance of lags in behavior or the result of decision making. This information can be significant because many economic policies can be expected to have an impact only after some time has passed.

Most books at the undergraduate level do not contain a discussion of econometric methods for panel data. However, economists now recognize that some questions are difficult, if not impossible, to answer satisfactorily without panel data. As you will see, we can make considerable progress with simple panel data analysis, a method that is not much more difficult than dealing with a standard cross-sectional data set.

1-3e A Comment on Data Structures

Part 1 of this text is concerned with the analysis of cross-sectional data, because this poses the fewest conceptual and technical difficulties. At the same time, it illustrates most of the key themes of econometric analysis. We will use the methods and insights from cross-sectional analysis in the remainder of the text.

Although the econometric analysis of time series uses many of the same tools as cross-sectional analysis, it is more complicated because of the trending, highly persistent nature of many economic time series. Examples that have been traditionally used to illustrate the manner in which econometric methods can be applied to time series data are now widely believed to be flawed. It makes little sense to use such examples initially, because this practice will only reinforce poor econometric practice. Therefore, we will postpone the treatment of time series econometrics until Part 2, when the important issues concerning trends, persistence, dynamics, and seasonality will be introduced.

In Part 3, we will treat pooled cross sections and panel data explicitly. The analysis of independently pooled cross sections and simple panel data analysis are fairly straightforward extensions of pure cross-sectional analysis. Nevertheless, we will wait until Chapter 13 to deal with these topics.

1-4 Causality, Ceteris Paribus, and Counterfactual Reasoning

In most tests of economic theory, and certainly for evaluating public policy, the economist's goal is to infer that one variable (such as education) has a **causal effect** on another variable (such as worker productivity). Simply finding an association between two or more variables might be suggestive, but unless causality can be established, it is rarely compelling.

The notion of **ceteris paribus**—which means "other (relevant) factors being equal"—plays an important role in causal analysis. This idea has been implicit in some of our earlier discussion, particularly Examples 1.1 and 1.2, but thus far we have not explicitly mentioned it.

You probably remember from introductory economics that most economic questions are ceteris paribus by nature. For example, in analyzing consumer demand, we are interested in knowing the effect of changing the price of a good on its quantity demanded, while holding all other factors—such as income, prices of other goods, and individual tastes—fixed. If other factors are not held fixed, then we cannot know the causal effect of a price change on quantity demanded.

Holding other factors fixed is critical for policy analysis as well. In the job training example (Example 1.2), we might be interested in the effect of another week of job training on wages, with all other components being equal (in particular, education and experience). If we succeed in holding all other relevant factors fixed and then find a link between job training and wages, we can conclude that job training has a causal effect on worker productivity. Although this may seem pretty simple, even at this early stage it should be clear that, except in very special cases, it will not be possible to literally hold all else equal. The key question in most empirical studies is: Have enough other factors been held fixed to make a case for causality? Rarely is an econometric study evaluated without raising this issue.

In most serious applications, the number of factors that can affect the variable of interest—such as criminal activity or wages—is immense, and the isolation of any particular variable may seem like a hopeless effort. However, we will eventually see that, when carefully applied, econometric methods can simulate a ceteris paribus experiment.

The notion of ceteris paribus also can be described through **counterfactual reasoning**, which has become an organizing theme in analyzing various interventions, such as policy changes. The idea is to imagine an economic unit, such as an individual or a firm, in two or more different states of the world. For example, consider studying the impact of a job training program on workers' earnings. For each worker in the relevant population, we can imagine what his or her subsequent earnings would be under two states of the world: having participated in the job training program and having not participated. By considering these **counterfactual outcomes** (also called *potential outcomes*), we easily "hold other factors fixed" because the counterfactual thought experiment applies to each individual separately. We can then think of causality as meaning that the outcome—in this case, labor earnings—in the two states of the world differs for at least some indiviuals. The fact that we will eventually observe each worker in only one state of the world raises important problems of estimation, but that is a separate issue from the issue of what we mean by causality. We formally introduce an apparatus for discussing counterfactual outcomes in Chapter 2.

At this point, we cannot yet explain how econometric methods can be used to estimate ceteris paribus effects, so we will consider some problems that can arise in trying to infer causality in economics. We do not use any equations in this discussion. Instead, in each example, we will discuss what other factors we would like to hold fixed, and sprinkle in some counterfactual reasoning. For each example, inferring causality becomes relatively easy if we could conduct an appropriate experiment. Thus, it is useful to describe how such an experiment might be structured, and to observe that, in most cases, obtaining experimental data is impractical. It is also helpful to think about why the available data fail to have the important features of an experimental data set.

We rely, for now, on your intuitive understanding of such terms as *random*, *independence*, and *correlation*, all of which should be familiar from an introductory probability and statistics course. (These concepts are reviewed in Math Refresher B.) We begin with an example that illustrates some of these important issues.

EXAMPLE 1.3 **Effects of Fertilizer on Crop Yield**

Some early econometric studies [for example, Griliches (1957)] considered the effects of new fertilizers on crop yields. Suppose the crop under consideration is soybeans. Because fertilizer amount is only one factor affecting yields—some others include rainfall, quality of land, and presence of parasites—this issue must be posed as a ceteris paribus question. One way to determine the causal effect of fertilizer amount on soybean yield is to conduct an experiment, which might include the following steps. Choose several one-acre plots of land. Apply different amounts of fertilizer to each plot and subsequently measure the yields; this gives us a cross-sectional data set. Then, use statistical methods (to be introduced in Chapter 2) to measure the association between yields and fertilizer amounts.

As described earlier, this may not seem like a very good experiment because we have said nothing about choosing plots of land that are identical in all respects except for the amount of fertilizer. In fact, choosing plots of land with this feature is not feasible: some of the factors, such as land quality, cannot even be fully observed. How do we know the results of this experiment can be used to measure the ceteris paribus effect of fertilizer? The answer depends on the specifics of how fertilizer amounts are chosen. If the levels of fertilizer are assigned to plots independently of other plot features that affect yield—that is, other characteristics of plots are completely ignored when deciding on fertilizer amounts—then we are in business. We will justify this statement in Chapter 2.

The next example is more representative of the difficulties that arise when inferring causality in applied economics.

EXAMPLE 1.4 **Measuring the Return to Education**

Labor economists and policy makers have long been interested in the "return to education." Somewhat informally, the question is posed as follows: If a person is chosen from the population and given another year of education, by how much will his or her wage increase? As with the previous examples, this is a ceteris paribus question, which implies that all other factors are held fixed while another year of education is given to the person. Notice the element of counterfactual reasoning here: we can imagine the wage of each individual varying with different levels of education, that is, in different states of the world. Eventually, we obtain data on each worker in only one state of the world: the education level they actually wound up with, through perhaps a complicated process of intellectual ability, motivation for learning, parental input, and societal influences.

We can imagine a social planner designing an experiment to get at this issue, much as the agricultural researcher can design an experiment to estimate fertilizer effects. Assume, for the moment, that the social planner has the ability to assign any level of education to any person. How would this planner emulate the fertilizer experiment in Example 1.3? The planner would choose a group of people and randomly assign each person an amount of education; some people are given an eighth-grade education, some are given a high school education, some are given two years of college, and so on. Subsequently, the planner measures wages for this group of people (where we assume that each person then works in a job). The people here are like the plots in the fertilizer example, where education plays the role of fertilizer and wage rate plays the role of soybean yield. As with Example 1.3, if levels of education are assigned independently of other characteristics that affect productivity (such as experience and innate ability), then an analysis that ignores these other factors will yield useful results. Again, it will take some effort in Chapter 2 to justify this claim; for now, we state it without support.

Unlike the fertilizer-yield example, the experiment described in Example 1.4 is unfeasible. The ethical issues, not to mention the economic costs, associated with randomly determining education levels for a group of individuals are obvious. As a logistical matter, we could not give someone only an eighth-grade education if he or she already has a college degree.

Even though experimental data cannot be obtained for measuring the return to education, we can certainly collect nonexperimental data on education levels and wages for a large group by sampling randomly from the population of working people. Such data are available from a variety of surveys used in labor economics, but these data sets have a feature that makes it difficult to estimate the ceteris paribus return to education. People *choose* their own levels of education; therefore, education levels are probably not determined independently of all other factors affecting wage. This problem is a feature shared by most nonexperimental data sets.

One factor that affects wage is experience in the workforce. Because pursuing more education generally requires postponing entering the workforce, those with more education usually have less experience. Thus, in a nonexperimental data set on wages and education, education is likely to be negatively associated with a key variable that also affects wage. It is also believed that people with more innate ability often choose higher levels of education. Because higher ability leads to higher wages, we again have a correlation between education and a critical factor that affects wage.

The omitted factors of experience and ability in the wage example have analogs in the fertilizer example. Experience is generally easy to measure and therefore is similar to a variable such as rainfall. Ability, on the other hand, is nebulous and difficult to quantify; it is similar to land quality in the fertilizer example. As we will see throughout this text, accounting for other observed factors, such as experience, when estimating the ceteris paribus effect of another variable, such as education, is relatively straightforward. We will also find that accounting for inherently unobservable factors, such as ability, is much more problematic. It is fair to say that many of the advances in econometric methods have tried to deal with unobserved factors in econometric models.

One final parallel can be drawn between Examples 1.3 and 1.4. Suppose that in the fertilizer example, the fertilizer amounts were not entirely determined at random. Instead, the assistant who

chose the fertilizer levels thought it would be better to put more fertilizer on the higher-quality plots of land. (Agricultural researchers should have a rough idea about which plots of land are of better quality, even though they may not be able to fully quantify the differences.) This situation is completely analogous to the level of schooling being related to unobserved ability in Example 1.4. Because better land leads to higher yields, and more fertilizer was used on the better plots, any observed relationship between yield and fertilizer might be spurious.

Difficulty in inferring causality can also arise when studying data at fairly high levels of aggregation, as the next example on city crime rates shows.

EXAMPLE 1.5 The Effect of Law Enforcement on City Crime Levels

The issue of how best to prevent crime has been, and will probably continue to be, with us for some time. One especially important question in this regard is: Does the presence of more police officers on the street deter crime?

The ceteris paribus question is easy to state: If a city is randomly chosen and given, say, ten additional police officers, by how much would its crime rates fall? Closely related to this thought experiment is explicitly setting up counterfactual outcomes: For a given city, what would its crime rate be under varying sizes of the police force? Another way to state the question is: If two cities are the same in all respects, except that city A has ten more police officers than city B, by how much would the two cities' crime rates differ?

It would be virtually impossible to find pairs of communities identical in all respects except for the size of their police force. Fortunately, econometric analysis does not require this. What we do need to know is whether the data we can collect on community crime levels and the size of the police force can be viewed as experimental. We can certainly imagine a true experiment involving a large collection of cities where we dictate how many police officers each city will use for the upcoming year.

Although policies can be used to affect the size of police forces, we clearly cannot tell each city how many police officers it can hire. If, as is likely, a city's decision on how many police officers to hire is correlated with other city factors that affect crime, then the data must be viewed as nonexperimental. In fact, one way to view this problem is to see that a city's choice of police force size and the amount of crime are *simultaneously determined*. We will explicitly address such problems in Chapter 16.

The first three examples we have discussed have dealt with cross-sectional data at various levels of aggregation (for example, at the individual or city levels). The same hurdles arise when inferring causality in time series problems.

EXAMPLE 1.6 The Effect of the Minimum Wage on Unemployment

An important, and perhaps contentious, policy issue concerns the effect of the minimum wage on unemployment rates for various groups of workers. Although this problem can be studied in a variety of data settings (cross-sectional, time series, or panel data), time series data are often used to look at aggregate effects. An example of a time series data set on unemployment rates and minimum wages was given in Table 1.3.

Standard supply and demand analysis implies that, as the minimum wage is increased above the market clearing wage, we slide up the demand curve for labor and total employment decreases. (Labor supply exceeds labor demand.) To quantify this effect, we can study the relationship between employment and the minimum wage over time. In addition to some special difficulties that can arise in dealing with time series data, there are possible problems with inferring causality. The minimum wage in the United States is not determined in a vacuum. Various economic and political forces impinge on the final minimum wage for any given year. (The minimum wage, once determined, is usually in place for several years, unless it is indexed for inflation.) Thus, it is probable that the amount of the minimum wage is related to other factors that have an effect on employment levels.

We can imagine the U.S. government conducting an experiment to determine the employment effects of the minimum wage (as opposed to worrying about the welfare of low-wage workers). The minimum wage could be randomly set by the government each year, and then the employment outcomes could be tabulated. The resulting experimental time series data could then be analyzed using fairly simple econometric methods. But this scenario hardly describes how minimum wages are set.

If we can control enough other factors relating to employment, then we can still hope to estimate the ceteris paribus effect of the minimum wage on employment. In this sense, the problem is very similar to the previous cross-sectional examples.

Even when economic theories are not most naturally described in terms of causality, they often have predictions that can be tested using econometric methods. The following example demonstrates this approach.

EXAMPLE 1.7 **The Expectations Hypothesis**

The *expectations hypothesis* from financial economics states that, given all information available to investors at the time of investing, the *expected* return on any two investments is the same. For example, consider two possible investments with a three-month investment horizon, purchased at the same time: (1) Buy a three-month T-bill with a face value of $10,000, for a price below $10,000; in three months, you receive $10,000. (2) Buy a six-month T-bill (at a price below $10,000) and, in three months, sell it as a three-month T-bill. Each investment requires roughly the same amount of initial capital, but there is an important difference. For the first investment, you know exactly what the return is at the time of purchase because you know the initial price of the three-month T-bill, along with its face value. This is not true for the second investment: although you know the price of a six-month T-bill when you purchase it, you do not know the price you can sell it for in three months. Therefore, there is uncertainty in this investment for someone who has a three-month investment horizon.

The actual returns on these two investments will usually be different. According to the expectations hypothesis, the expected return from the second investment, given all information at the time of investment, should equal the return from purchasing a three-month T-bill. This theory turns out to be fairly easy to test, as we will see in Chapter 11.

Summary

In this introductory chapter, we have discussed the purpose and scope of econometric analysis. Econometrics is used in all applied economics fields to test economic theories, to inform government and private policy makers, and to predict economic time series. Sometimes, an econometric model is derived from a formal economic model, but in other cases, econometric models are based on informal economic reasoning and intuition. The goals of any econometric analysis are to estimate the parameters in the model and to test hypotheses about these parameters; the values and signs of the parameters determine the validity of an economic theory and the effects of certain policies.

Cross-sectional, time series, pooled cross-sectional, and panel data are the most common types of data structures that are used in applied econometrics. Data sets involving a time dimension, such as time series and panel data, require special treatment because of the correlation across time of most economic time series. Other issues, such as trends and seasonality, arise in the analysis of time series data but not cross-sectional data.

In Section 1-4, we discussed the notions of causality, ceteris paribus, and counterfactuals. In most cases, hypotheses in the social sciences are ceteris paribus in nature: all other relevant factors must be fixed when studying the relationship between two variables. As we discussed, one way to think of the ceteris paribus requirement is to undertake a thought experiment where the same economic unit operates in different states of the world, such as different policy regimes. Because of the nonexperimental nature of most data collected in the social sciences, uncovering causal relationships is very challenging.

Key Terms

Causal Effect
Ceteris Paribus
Counterfactual Outcomes
Counterfactual Reasoning
Cross-Sectional Data Set
Data Frequency
Econometric Model
Economic Model
Empirical Analysis
Experimental Data
Nonexperimental Data
Observational Data
Panel Data
Pooled Cross Section
Random Sampling
Retrospective Data
Time Series Data

Problems

1 Suppose that you are asked to conduct a study to determine whether smaller class sizes lead to improved student performance of fourth graders.

(i) If you could conduct any experiment you want, what would you do? Be specific.

(ii) More realistically, suppose you can collect observational data on several thousand fourth graders in a given state. You can obtain the size of their fourth-grade class and a standardized test score taken at the end of fourth grade. Why might you expect a negative correlation between class size and test score?

(iii) Would a negative correlation necessarily show that smaller class sizes cause better performance? Explain.

2 A justification for job training programs is that they improve worker productivity. Suppose that you are asked to evaluate whether more job training makes workers more productive. However, rather than having data on individual workers, you have access to data on manufacturing firms in Ohio. In particular, for each firm, you have information on hours of job training per worker (*training*) and number of nondefective items produced per worker hour (*output*).

(i) Carefully state the ceteris paribus thought experiment underlying this policy question.

(ii) Does it seem likely that a firm's decision to train its workers will be independent of worker characteristics? What are some of those measurable and unmeasurable worker characteristics?

(iii) Name a factor other than worker characteristics that can affect worker productivity.

(iv) If you find a positive correlation between *output* and *training*, would you have convincingly established that job training makes workers more productive? Explain.

3 Suppose at your university you are asked to find the relationship between weekly hours spent studying (*study*) and weekly hours spent working (*work*). Does it make sense to characterize the problem as inferring whether *study* "causes" *work* or *work* "causes" *study*? Explain.

4 States (and provinces) that have control over taxation sometimes reduce taxes in an attempt to spur economic growth. Suppose that you are hired by a state to estimate the effect of corporate tax rates on, say, the growth in per capita gross state product (GSP).

(i) What kind of data would you need to collect to undertake a statistical analysis?

(ii) Is it feasible to do a controlled experiment? What would be required?

(iii) Is a correlation analysis between GSP growth and tax rates likely to be convincing? Explain.

Computer Exercises

C1 Use the data in WAGE1 for this exercise.

(i) Find the average education level in the sample. What are the lowest and highest years of education?

(ii) Find the average hourly wage in the sample. Does it seem high or low?

(iii) The wage data are reported in 1976 dollars. Using the Internet or a printed source, find the Consumer Price Index (CPI) for the years 1976 and 2013.

(iv) Use the CPI values from part (iii) to find the average hourly wage in 2013 dollars. Now does the average hourly wage seem reasonable?

(v) How many women are in the sample? How many men?

C2 Use the data in BWGHT to answer this question.

(i) How many women are in the sample, and how many report smoking during pregnancy?

(ii) What is the average number of cigarettes smoked per day? Is the average a good measure of the "typical" woman in this case? Explain.

(iii) Among women who smoked during pregnancy, what is the average number of cigarettes smoked per day? How does this compare with your answer from part (ii), and why?

(iv) Find the average of *fatheduc* in the sample. Why are only 1,192 observations used to compute this average?

(v) Report the average family income and its standard deviation in dollars.

C3 The data in MEAP01 are for the state of Michigan in the year 2001. Use these data to answer the following questions.

(i) Find the largest and smallest values of *math4*. Does the range make sense? Explain.

(ii) How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?

(iii) How many schools have math pass rates of exactly 50%?

(iv) Compare the average pass rates for the math and reading scores. Which test is harder to pass?

(v) Find the correlation between *math4* and *read4*. What do you conclude?

(vi) The variable *exppp* is expenditure per pupil. Find the average of *exppp* along with its standard deviation. Would you say there is wide variation in per pupil spending?

(vii) Suppose School A spends \$6,000 per student and School B spends \$5,500 per student. By what percentage does School A's spending exceed School B's? Compare this to $100 \cdot [\log(6{,}000) - \log(5{,}500)]$, which is the approximation percentage difference based on the difference in the natural logs. (See Section A.4 in Math Refresher A.)

C4 The data in JTRAIN2 come from a job training experiment conducted for low-income men during 1976–1977; see Lalonde (1986).

(i) Use the indicator variable *train* to determine the fraction of men receiving job training.

(ii) The variable *re78* is earnings from 1978, measured in thousands of 1982 dollars. Find the averages of *re78* for the sample of men receiving job training and the sample not receiving job training. Is the difference economically large?

(iii) The variable *unem78* is an indicator of whether a man is unemployed or not in 1978. What fraction of the men who received job training are unemployed? What about for men who did not receive job training? Comment on the difference.

(iv) From parts (ii) and (iii), does it appear that the job training program was effective? What would make our conclusions more convincing?

C5 The data in FERTIL2 were collected on women living in the Republic of Botswana in 1988. The variable *children* refers to the number of living children. The variable *electric* is a binary indicator equal to one if the woman's home has electricity, and zero if not.

(i) Find the smallest and largest values of *children* in the sample. What is the average of *children*?

(ii) What percentage of women have electricity in the home?

(iii) Compute the average of *children* for those without electricity and do the same for those with electricity. Comment on what you find.

(iv) From part (iii), can you infer that having electricity "causes" women to have fewer children? Explain.

C6 Use the data in COUNTYMURDERS to answer this question. Use only the year 1996. The variable *murders* is the number of murders reported in the county. The variable *execs* is the number of executions that took place of people sentenced to death in the given county. Most states in the United States have the death penalty, but several do not.

(i) How many counties are there in the data set? Of these, how many have zero murders? What percentage of counties have zero executions? (Remember, use only the 1996 data.)

(ii) What is the largest number of murders? What is the largest number of executions? Compute the average number of executions and explain why it is so small.

(iii) Compute the correlation coefficient between *murders* and *execs* and describe what you find.

(iv) You should have computed a positive correlation in part (iii). Do you think that more executions *cause* more murders to occur? What might explain the positive correlation?

C7 The data set in ALCOHOL contains information on a sample of men in the United States. Two key variables are self-reported employment status and alcohol abuse (along with many other variables). The variables *employ* and *abuse* are both binary, or indicator, variables: they take on only the values zero and one.

(i) What percentage of the men in the sample report abusing alcohol? What is the employment rate?

(ii) Consider the group of men who abuse alcohol. What is the employment rate?

(iii) What is the employment rate for the group of men who do not abuse alcohol?

(iv) Discuss the difference in your answers to parts (ii) and (iii). Does this allow you to conclude that alcohol abuse causes unemployment?

C8 The data in ECONMATH were obtained on students from a large university course in introductory microeconomics. For this problem, we are interested in two variables: *score*, which is the final course score, and *econhs*, which is a binary variable indicating whether a student took an economics course in high school.

(i) How many students are in the sample? How many students report taking an economics course in high school?

(ii) Find the average of *score* for those students who did take a high school economics class. How does it compare with the average of *score* for those who did not?

(iii) Do the findings in part (ii) necessarily tell you anything about the causal effect of taking high school economics on college course peformance? Explain.

(iv) If you want to obtain a good causal estimate of the effect of taking a high school economics course using the difference in averages, what experiment would you run?

PART 1

Regression Analysis with Cross-Sectional Data

Part 1 of the text covers regression analysis with cross-sectional data. It builds upon a solid base of college algebra and basic concepts in probability and statistics. Math Refresher A, B, and C contain complete reviews of these topics.

Chapter 2 begins with the simple linear regression model, where we explain one variable in terms of another variable. Although simple regression is not widely used in applied econometrics, it is used occasionally and serves as a natural starting point because the algebra and interpretations are relatively straightforward.

Chapters 3 and 4 cover the fundamentals of multiple regression analysis, where we allow more than one variable to affect the variable we are trying to explain. Multiple regression is still the most commonly used method in empirical research, and so these chapters deserve careful attention. Chapter 3 focuses on the algebra of the method of ordinary least squares (OLS), while also establishing conditions under which the OLS estimator is unbiased and best linear unbiased. Chapter 4 covers the important topic of statistical inference.

Chapter 5 discusses the large sample, or asymptotic, properties of the OLS estimators. This provides justification of the inference procedures in Chapter 4 when the errors in a regression model are not normally distributed. Chapter 6 covers some additional topics in regression analysis, including advanced functional form issues, data scaling, prediction, and goodness-of-fit. Chapter 7 explains how qualitative information can be incorporated into multiple regression models.

Chapter 8 illustrates how to test for and correct the problem of heteroskedasticity, or nonconstant variance, in the error terms. We show how the usual OLS statistics can be adjusted, and we also present an extension of OLS, known as *weighted least squares*, which explicitly accounts for different variances in the errors. Chapter 9 delves further into the very important problem of correlation between the error term and one or more of the explanatory variables. We demonstrate how the availability of a proxy variable can solve the omitted variables problem. In addition, we establish the bias and inconsistency in the OLS estimators in the presence of certain kinds of measurement errors in the variables. Various data problems are also discussed, including the problem of outliers.

CHAPTER 2

The Simple Regression Model

The simple regression model can be used to study the relationship between two variables. For reasons we will see, the simple regression model has limitations as a general tool for empirical analysis. Nevertheless, it is sometimes appropriate as an empirical tool. Learning how to interpret the simple regression model is good practice for studying multiple regression, which we will do in subsequent chapters.

2-1 Definition of the Simple Regression Model

Much of applied econometric analysis begins with the following premise: y and x are two variables, representing some population, and we are interested in "explaining y in terms of x," or in "studying how y varies with changes in x." We discussed some examples in Chapter 1, including: y is soybean crop yield and x is amount of fertilizer; y is hourly wage and x is years of education; and y is a community crime rate and x is number of police officers.

In writing down a model that will "explain y in terms of x," we must confront three issues. First, because there is never an exact relationship between two variables, how do we allow for other factors to affect y? Second, what is the functional relationship between y and x? And third, how can we be sure we are capturing a ceteris paribus relationship between y and x (if that is a desired goal)?

We can resolve these ambiguities by writing down an equation relating y to x. A simple equation is

$$y = \beta_0 + \beta_1 x + u. \qquad [2.1]$$

Equation (2.1), which is assumed to hold in the population of interest, defines the **simple linear regression model**. It is also called the *two-variable linear regression model* or *bivariate linear*

TABLE 2.1 Terminology for Simple Regression

Y	*X*
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

regression model because it relates the two variables x and y. We now discuss the meaning of each of the quantities in equation (2.1). [Incidentally, the term "regression" has origins that are not especially important for most modern econometric applications, so we will not explain it here. See Stigler (1986) for an engaging history of regression analysis.]

When related by equation (2.1), the variables y and x have several different names used interchangeably, as follows: y is called the **dependent variable**, the **explained variable**, the **response variable**, the **predicted variable**, or the **regressand**; x is called the **independent variable**, the **explanatory variable**, the **control variable**, the **predictor variable**, or the **regressor**. (The term **covariate** is also used for x.) The terms "dependent variable" and "independent variable" are frequently used in econometrics. But be aware that the label "independent" here does not refer to the statistical notion of independence between random variables (see Math Refresher B).

The terms "explained" and "explanatory" variables are probably the most descriptive. "Response" and "control" are used mostly in the experimental sciences, where the variable x is under the experimenter's control. We will not use the terms "predicted variable" and "predictor," although you sometimes see these in applications that are purely about prediction and not causality. Our terminology for simple regression is summarized in Table 2.1.

The variable u, called the **error term** or **disturbance** in the relationship, represents factors other than x that affect y. A simple regression analysis effectively treats all factors affecting y other than x as being unobserved. You can usefully think of u as standing for "unobserved."

Equation (2.1) also addresses the issue of the functional relationship between y and x. If the other factors in u are held fixed, so that the change in u is zero, $\Delta u = 0$, then x has a *linear* effect on y:

$$\Delta y = \beta_1 \Delta x \text{ if } \Delta u = 0. \qquad [2.2]$$

Thus, the change in y is simply β_1 multiplied by the change in x. This means that β_1 is the **slope parameter** in the relationship between y and x, holding the other factors in u fixed; it is of primary interest in applied economics. The **intercept parameter** β_0, sometimes called the *constant term*, also has its uses, although it is rarely central to an analysis.

EXAMPLE 2.1 Soybean Yield and Fertilizer

Suppose that soybean yield is determined by the model

$$yield = \beta_0 + \beta_1 fertilizer + u, \qquad [2.3]$$

so that $y = yield$ and $x = fertilizer$. The agricultural researcher is interested in the effect of fertilizer on yield, holding other factors fixed. This effect is given by β_1. The error term u contains factors such as land quality, rainfall, and so on. The coefficient β_1 measures the effect of fertilizer on yield, holding other factors fixed: $\Delta yield = \beta_1 \Delta\, fertilizer$.

EXAMPLE 2.2 **A Simple Wage Equation**

A model relating a person's wage to observed education and other unobserved factors is

$$wage = \beta_0 + \beta_1 educ + u. \quad [2.4]$$

If *wage* is measured in dollars per hour and *educ* is years of education, then β_1 measures the change in hourly wage given another year of education, holding all other factors fixed. Some of those factors include labor force experience, innate ability, tenure with current employer, work ethic, and numerous other things.

The linearity of equation (2.1) implies that a one-unit change in x has the *same* effect on y, regardless of the initial value of x. This is unrealistic for many economic applications. For example, in the wage-education example, we might want to allow for *increasing* returns: the next year of education has a *larger* effect on wages than did the previous year. We will see how to allow for such possibilities in Section 2-4.

The most difficult issue to address is whether model (2.1) really allows us to draw ceteris paribus conclusions about how x affects y. We just saw in equation (2.2) that β_1 *does* measure the effect of x on y, holding all other factors (in u) fixed. Is this the end of the causality issue? Unfortunately, no. How can we hope to learn in general about the ceteris paribus effect of x on y, holding other factors fixed, when we are ignoring all those other factors?

Section 2-5 will show that we are only able to get reliable estimators of β_0 and β_1 from a random sample of data when we make an assumption restricting how the unobservable u is related to the explanatory variable x. Without such a restriction, we will not be able to estimate the ceteris paribus effect, β_1. Because u and x are random variables, we need a concept grounded in probability.

Before we state the key assumption about how x and u are related, we can always make one assumption about u. As long as the intercept β_0 is included in the equation, nothing is lost by assuming that the average value of u in the population is zero. Mathematically,

$$E(u) = 0. \quad [2.5]$$

Assumption (2.5) says nothing about the relationship between u and x, but simply makes a statement about the distribution of the unobserved factors in the population. Using the previous examples for illustration, we can see that assumption (2.5) is not very restrictive. In Example 2.1, we lose nothing by normalizing the unobserved factors affecting soybean yield, such as land quality, to have an average of zero in the population of all cultivated plots. The same is true of the unobserved factors in Example 2.2. Without loss of generality, we can assume that things such as average ability are zero in the population of all working people. If you are not convinced, you should work through Problem 2 to see that we can always redefine the intercept in equation (2.1) to make equation (2.5) true.

We now turn to the crucial assumption regarding how u and x are related. A natural measure of the association between two random variables is the *correlation coefficient*. (See Math Refresher B for definition and properties.) If u and x are *uncorrelated*, then, as random variables, they are not *linearly* related. Assuming that u and x are uncorrelated goes a long way toward defining the sense in which u and x should be unrelated in equation (2.1). But it does not go far enough, because correlation measures only linear dependence between u and x. Correlation has a somewhat counterintuitive feature: it is possible for u to be uncorrelated with x while being correlated with functions of x, such as x^2. (See Section B-4 in Math Refresher B for further discussion.) This possibility is not acceptable for most regression purposes, as it causes problems for interpreting the model and for deriving statistical properties. A better assumption involves the *expected value of u given x*.

Because u and x are random variables, we can define the conditional distribution of u given any value of x. In particular, for any x, we can obtain the expected (or average) value of u for that slice of

the population described by the value of x. The crucial assumption is that the average value of u does *not* depend on the value of x. We can write this assumption as

$$\mathrm{E}(u|x) = \mathrm{E}(u). \qquad [2.6]$$

Equation (2.6) says that the average value of the unobservables is the same across all slices of the population determined by the value of x and that the common average is necessarily equal to the average of u over the entire population. When assumption (2.6) holds, we say that u is **mean independent** of x. (Of course, mean independence is implied by full independence between u and x, an assumption often used in basic probability and statistics.) When we combine mean independence with assumption (2.5), we obtain the **zero conditional mean assumption**, $\mathrm{E}(u|x) = 0$. It is critical to remember that equation (2.6) is the assumption with impact; assumption (2.5) essentially defines the intercept, β_0.

Let us see what equation (2.6) entails in the wage example. To simplify the discussion, assume that u is the same as innate ability. Then equation (2.6) requires that the average level of ability is the same, regardless of years of education. For example, if $\mathrm{E}(abil|8)$ denotes the average ability for the group of all people with eight years of education, and $\mathrm{E}(abil|16)$ denotes the average ability among people in the population with sixteen years of education, then equation (2.6) implies that these must be the same. In fact, the average ability level must be the same for *all* education levels. If, for example, we think that average ability increases with years of education, then equation (2.6) is false. (This would happen if, on average, people with more ability choose to become more educated.) As we cannot observe innate ability, we have no way of knowing whether or not average ability is the same for all education levels. But this is an issue that we must address before relying on simple regression analysis.

GOING FURTHER 2.1

Suppose that a score on a final exam, *score*, depends on classes attended (*attend*) and unobserved factors that affect exam performance (such as student ability). Then

$$score = \beta_0 + \beta_1 attend + u. \qquad [2.7]$$

When would you expect this model to satisfy equation (2.6)?

In the fertilizer example, if fertilizer amounts are chosen independently of other features of the plots, then equation (2.6) will hold: the average land quality will not depend on the amount of fertilizer. However, if more fertilizer is put on the higher-quality plots of land, then the expected value of u changes with the level of fertilizer, and equation (2.6) fails.

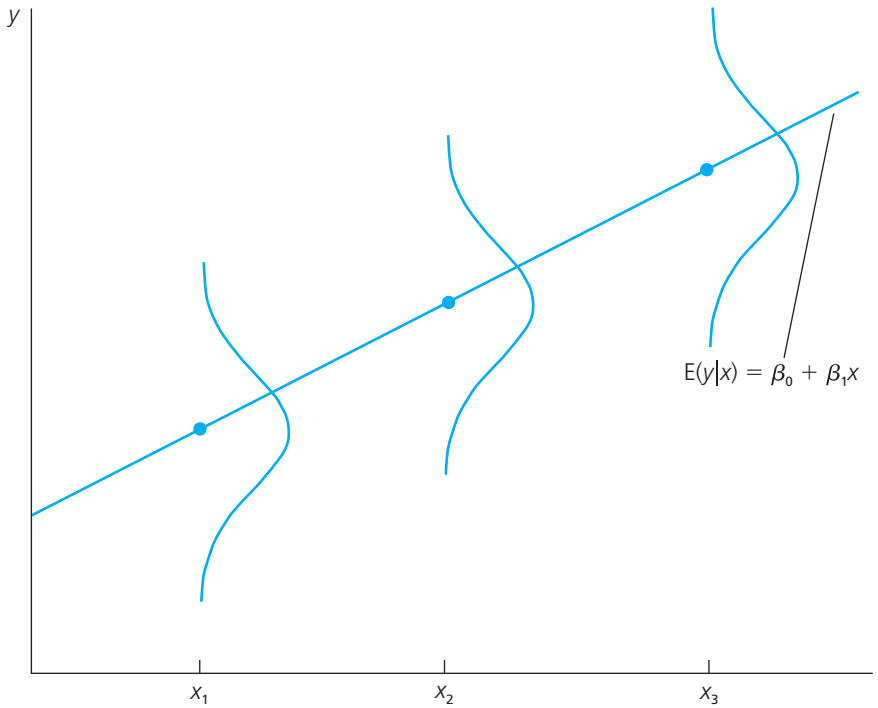
The zero conditional mean assumption gives β_1 another interpretation that is often useful. Taking the expected value of equation (2.1) conditional on x and using $E(u|x) = 0$ gives

$$\mathrm{E}(y|x) = \beta_0 + \beta_1 x. \qquad [2.8]$$

Equation (2.8) shows that the **population regression function** (**PRF**), $\mathrm{E}(y|x)$, is a linear function of x. The linearity means that a one-unit increase in x changes the *expected value* of y by the amount β_1. For any given value of x, the distribution of y is centered about $\mathrm{E}(y|x)$, as illustrated in Figure 2.1.

It is important to understand that equation (2.8) tells us how the *average* value of y changes with x; it does not say that y equals $\beta_0 + \beta_1 x$ for all units in the population. For example, suppose that x is the high school grade point average and y is the college GPA, and we happen to know that $\mathrm{E}(colGPA|hsGPA) = 1.5 + 0.5\, hsGPA$. [Of course, in practice, we never know the population intercept and slope, but it is useful to pretend momentarily that we do to understand the nature of equation (2.8).] This GPA equation tells us the *average* college GPA among all students who have a given high school GPA. So suppose that $hsGPA = 3.6$. Then the average *colGPA* for all high school graduates who attend college with $hsGPA = 3.6$ is $1.5 + 0.5(3.6) = 3.3$. We are certainly *not* saying that every student with $hsGPA = 3.6$ will have a 3.3 college GPA; this is clearly false. The PRF gives us a relationship between the average level of y at different levels of x. Some students with $hsGPA = 3.6$ will have a college GPA higher than 3.3, and some will have a lower college GPA. Whether the actual *colGPA* is above or below 3.3 depends on the unobserved factors in u, and those differ among students even within the slice of the population with $hsGPA = 3.6$.

FIGURE 2.1 $E(y|x)$ as a linear function of x.



Given the zero conditional mean assumption $E(u|x) = 0$, it is useful to view equation (2.1) as breaking y into two components. The piece $\beta_0 + \beta_1 x$, which represents $E(y|x)$, is called the *systematic part* of y—that is, the part of y explained by x—and u is called the *unsystematic part*, or the part of y not explained by x. In Chapter 3, when we introduce more than one explanatory variable, we will discuss how to determine how large the systematic part is relative to the unsystematic part.

In the next section, we will use assumptions (2.5) and (2.6) to motivate estimators of β_0 and β_1 given a random sample of data. The zero conditional mean assumption also plays a crucial role in the statistical analysis in Section 2-5.

2-2 Deriving the Ordinary Least Squares Estimates

Now that we have discussed the basic ingredients of the simple regression model, we will address the important issue of how to estimate the parameters β_0 and β_1 in equation (2.1). To do this, we need a sample from the population. Let $\{(x_i, y_i): i = 1, \ldots, n\}$ denote a random sample of size n from the population. Because these data come from equation (2.1), we can write

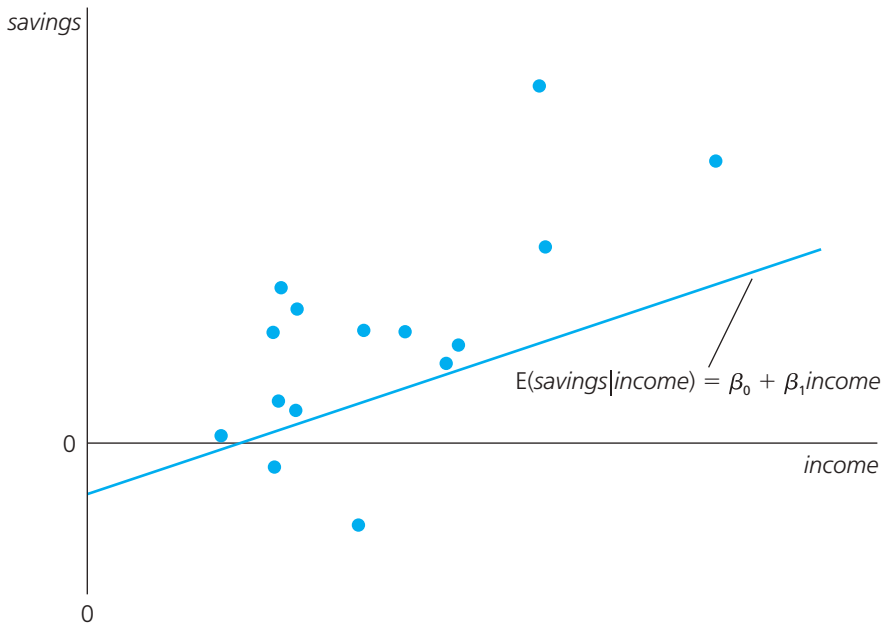
$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad [2.9]$$

for each i. Here, u_i is the error term for observation i because it contains all factors affecting y_i other than x_i.

As an example, x_i might be the annual income and y_i the annual savings for family i during a particular year. If we have collected data on 15 families, then $n = 15$. A scatterplot of such a data set is given in Figure 2.2, along with the (necessarily fictitious) population regression function.

We must decide how to use these data to obtain estimates of the intercept and slope in the population regression of savings on income.

FIGURE 2.2 Scatterplot of savings and income for 15 families, and the population regression $E(savings|income) = \beta_0 + \beta_1 income$.



There are several ways to motivate the following estimation procedure. We will use equation (2.5) and an important implication of assumption (2.6): in the population, u is uncorrelated with x. Therefore, we see that u has zero expected value and that the *covariance* between x and u is zero:

$$E(u) = 0 \qquad [2.10]$$

and

$$Cov(x,u) = E(xu) = 0, \qquad [2.11]$$

where the first equality in equation (2.11) follows from (2.10). (See Section B-4 in Math Refresher B for the definition and properties of covariance.) In terms of the observable variables x and y and the unknown parameters β_0 and β_1, equations (2.10) and (2.11) can be written as

$$E(y - \beta_0 - \beta_1 x) = 0 \qquad [2.12]$$

and

$$E[x(y - \beta_0 - \beta_1 x)] = 0, \qquad [2.13]$$

respectively. Equations (2.12) and (2.13) imply two restrictions on the joint probability distribution of (x,y) in the population. Because there are two unknown parameters to estimate, we might hope that equations (2.12) and (2.13) can be used to obtain good estimators of β_0 and β_1. In fact, they can be. Given a sample of data, we choose estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the *sample* counterparts of equations (2.12) and (2.13):

$$n^{-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad [2.14]$$

and

$$n^{-1}\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \qquad [2.15]$$

This is an example of the *method of moments* approach to estimation. (See Section C-4 in Math Refresher C for a discussion of different estimation approaches.) These equations can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Using the basic properties of the summation operator from Math Refresher A, equation (2.14) can be rewritten as

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}, \tag{2.16}$$

where $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ is the sample average of the y_i and likewise for $\bar{x}$. This equation allows us to write $\hat{\beta}_0$ in terms of $\hat{\beta}_1$, $\bar{y}$, and $\bar{x}$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2.17}$$

Therefore, once we have the slope estimate $\hat{\beta}_1$, it is straightforward to obtain the intercept estimate $\hat{\beta}_0$, given $\bar{y}$ and $\bar{x}$.

Dropping the n^{-1} in (2.15) (because it does not affect the solution) and plugging (2.17) into (2.15) yields

$$\sum_{i=1}^{n} x_i [y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] = 0,$$

which, upon rearrangement, gives

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}).$$

From basic properties of the summation operator [see (A-7) and (A-8) in Math Refresher A],

$$\sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ and } \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Therefore, provided that

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0, \tag{2.18}$$

the estimated slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{2.19}$$

Equation (2.19) is simply the sample covariance between x_i and y_i divided by the sample variance of x_i. Using simple algebra we can also write $\hat{\beta}_1$ as

$$\hat{\beta}_1 = \hat{\rho}_{xy} \cdot \left(\frac{\hat{\sigma}_y}{\hat{\sigma}_x} \right),$$

where $\hat{\rho}_{xy}$ is the sample correlation between x_i and y_i and $\hat{\sigma}_x$, $\hat{\sigma}_y$ denote the sample standard deviations. (See Math Refresher C for definitions of correlation and standard deviation. Dividing all sums by $n - 1$ does not affect the formulas.) An immediate implication is that if x_i and y_i are positively correlated in the sample then $\hat{\beta}_1 > 0$; if x_i and y_i are negatively correlated then $\hat{\beta}_1 < 0$.

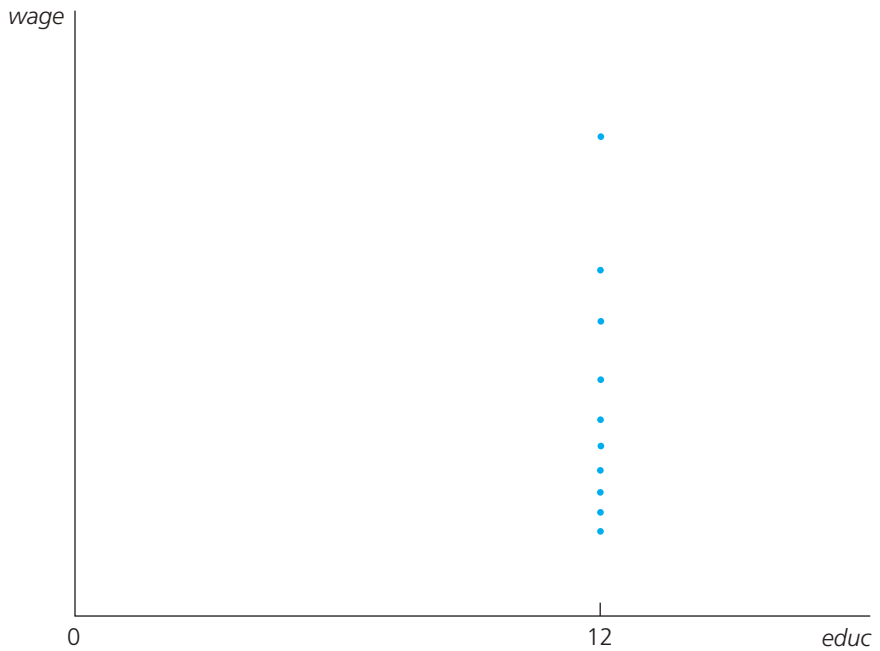
Not surprisingly, the formula for $\hat{\beta}_1$ in terms of the sample correlation and sample standard deviations is the sample analog of the population relationship

$$\beta_1 = \rho_{xy} \cdot \left(\frac{\sigma_y}{\sigma_x} \right),$$

where all quantities are defined for the entire population. Recognition that β_1 is just a scaled version of ρ_{xy} highlights an important limitation of simple regression when we do not have experimental data: in effect, simple regression is an analysis of correlation between two variables, and so one must be careful in inferring causality.

Although the method for obtaining (2.17) and (2.19) is motivated by (2.6), the only assumption needed to compute the estimates for a particular sample is (2.18). This is hardly an assumption at all: (2.18) is true, provided the x_i in the sample are not all equal to the same value. If (2.18) fails, then

FIGURE 2.3 A scatterplot of wage against education when $educ_i = 12$ for all i.



we have either been unlucky in obtaining our sample from the population or we have not specified an interesting problem (x does not vary in the population). For example, if $y = wage$ and $x = educ$, then (2.18) fails only if everyone in the sample has the same amount of education (for example, if everyone is a high school graduate; see Figure 2.3). If just one person has a different amount of education, then (2.18) holds, and the estimates can be computed.

The estimates given in (2.17) and (2.19) are called the **ordinary least squares (OLS)** estimates of β_0 and β_1. To justify this name, for any $\hat{\beta}_0$ and $\hat{\beta}_1$ define a **fitted value** for y when $x = x_i$ as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \qquad [2.20]$$

This is the value we predict for y when $x = x_i$ for the given intercept and slope. There is a fitted value for each observation in the sample. The **residual** for observation i is the difference between the actual y_i and its fitted value:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i. \qquad [2.21]$$

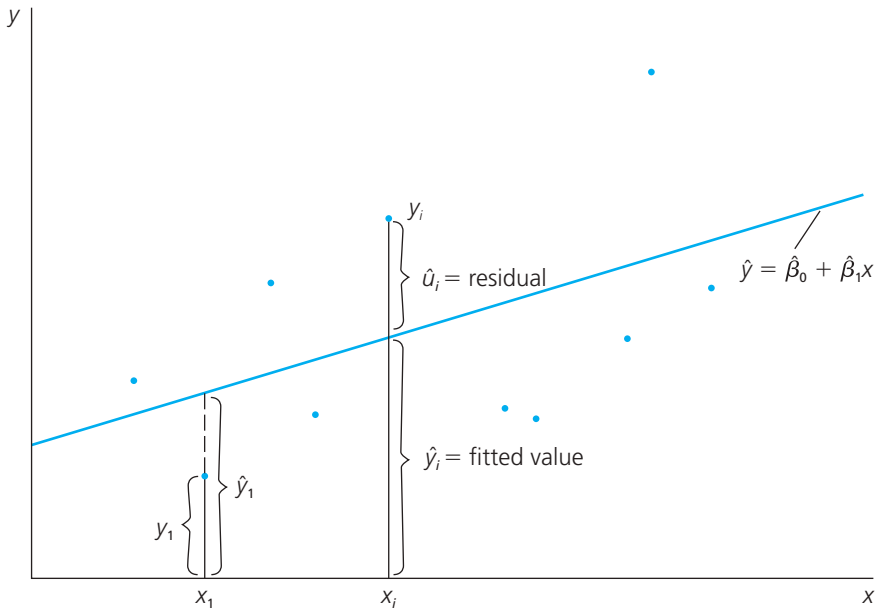
Again, there are n such residuals. [These are *not* the same as the errors in (2.9), a point we return to in Section 2-5.] The fitted values and residuals are indicated in Figure 2.4.

Now, suppose we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the **sum of squared residuals**,

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2, \qquad [2.22]$$

as small as possible. The appendix to this chapter shows that the conditions necessary for $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize (2.22) are given exactly by equations (2.14) and (2.15), without n^{-1}. Equations (2.14) and (2.15) are often called the **first order conditions** for the OLS estimates, a term that comes from optimization using calculus (see Math Refresher A). From our previous calculations, we know that the solutions to the OLS first order conditions are given by (2.17) and (2.19). The name "ordinary least squares" comes from the fact that these estimates minimize the sum of squared residuals.

FIGURE 2.4 Fitted values and residuals.



When we view ordinary least squares as minimizing the sum of squared residuals, it is natural to ask: why not minimize some other function of the residuals, such as the absolute values of the residuals? In fact, as we will discuss in the more advanced Section 9-6, minimizing the sum of the absolute values of the residuals is sometimes very useful. But it does have some drawbacks. First, we cannot obtain formulas for the resulting estimators; given a data set, the estimates must be obtained by numerical optimization routines. As a consequence, the statistical theory for estimators that minimize the sum of the absolute residuals is very complicated. Minimizing other functions of the residuals, say, the sum of the residuals each raised to the fourth power, has similar drawbacks. (We would never choose our estimates to minimize, say, the sum of the residuals themselves, as residuals large in magnitude but with opposite signs would tend to cancel out.) With OLS, we will be able to derive unbiasedness, consistency, and other important statistical properties relatively easily. Plus, as the motivation in equations (2.12) and (2.13) suggests, and as we will see in Section 2-5, OLS is suited for estimating the parameters appearing in the conditional mean function (2.8).

Once we have determined the OLS intercept and slope estimates, we form the **OLS regression line**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \qquad [2.23]$$

where it is understood that $\hat{\beta}_0$ and $\hat{\beta}_1$ have been obtained using equations (2.17) and (2.19). The notation $\hat{y}$, read as "y hat," emphasizes that the predicted values from equation (2.23) are estimates. The intercept, $\hat{\beta}_0$, is the predicted value of y when $x = 0$, although in some cases it will not make sense to set $x = 0$. In those situations, $\hat{\beta}_0$ is not, in itself, very interesting. When using (2.23) to compute predicted values of y for various values of x, we must account for the intercept in the calculations. Equation (2.23) is also called the **sample regression function (SRF)** because it is the estimated version of the population regression function $\mathrm{E}(y|x) = \beta_0 + \beta_1 x$. It is important to remember that the PRF is something fixed, but unknown, in the population. Because the SRF is obtained for a given sample of data, a new sample will generate a different slope and intercept in equation (2.23).

In most cases, the slope estimate, which we can write as

$$\hat{\beta}_1 = \Delta\hat{y}/\Delta x, \qquad [2.24]$$

is of primary interest. It tells us the amount by which $\hat{y}$ changes when x increases by one unit. Equivalently,

$$\Delta\hat{y} = \hat{\beta}_1\Delta x, \qquad [2.25]$$

so that given any change in x (whether positive or negative), we can compute the predicted change in y.

We now present several examples of simple regression obtained by using real data. In other words, we find the intercept and slope estimates with equations (2.17) and (2.19). Because these examples involve many observations, the calculations were done using an econometrics software package. At this point, you should be careful not to read too much into these regressions; they are not necessarily uncovering a causal relationship. We have said nothing so far about the statistical properties of OLS. In Section 2-5, we consider statistical properties after we explicitly impose assumptions on the population model equation (2.1).

EXAMPLE 2.3 CEO Salary and Return on Equity

For the population of chief executive officers, let y be annual salary (*salary*) in thousands of dollars. Thus, $y = 856.3$ indicates an annual salary of \$856,300, and $y = 1{,}452.6$ indicates a salary of \$1,452,600. Let x be the average return on equity (*roe*) for the CEO's firm for the previous three years. (Return on equity is defined in terms of net income as a percentage of common equity.) For example, if $roe = 10$, then average return on equity is 10%.

To study the relationship between this measure of firm performance and CEO compensation, we postulate the simple model

$$salary = \beta_0 + \beta_1 roe + u.$$

The slope parameter β_1 measures the change in annual salary, in thousands of dollars, when return on equity increases by one percentage point. Because a higher *roe* is good for the company, we think $\beta_1 > 0$.

The data set CEOSAL1 contains information on 209 CEOs for the year 1990; these data were obtained from *Business Week* (5/6/91). In this sample, the average annual salary is \$1,281,120, with the smallest and largest being \$223,000 and \$14,822,000, respectively. The average return on equity for the years 1988, 1989, and 1990 is 17.18%, with the smallest and largest values being 0.5% and 56.3%, respectively.

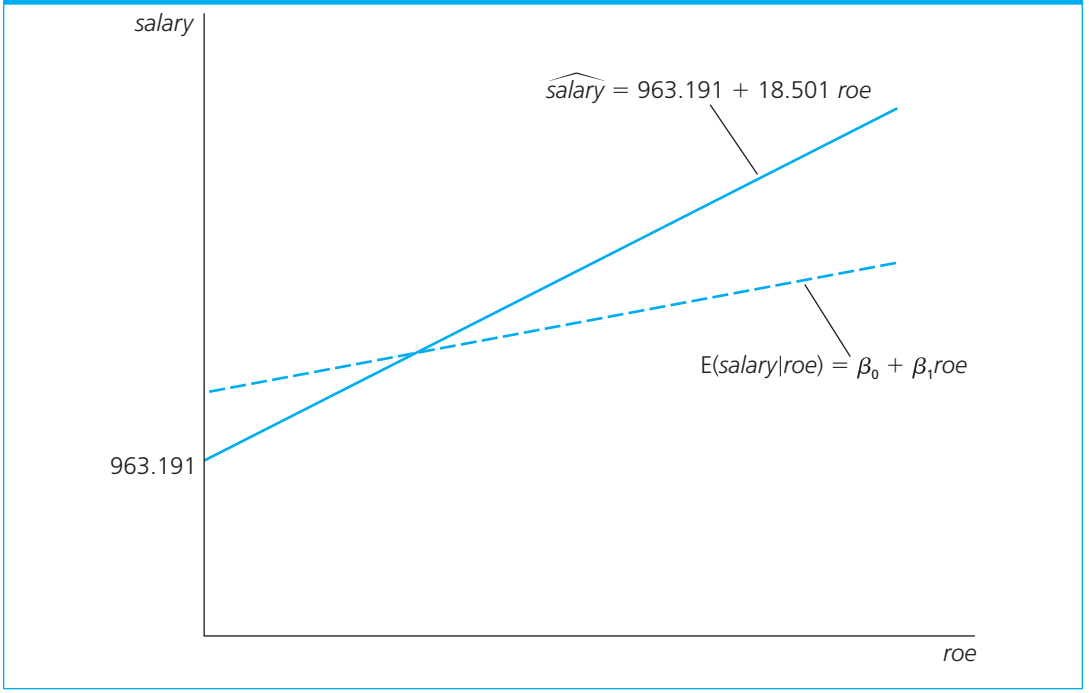
Using the data in CEOSAL1, the OLS regression line relating *salary* to *roe* is

$$\widehat{salary} = 963.191 + 18.501\, roe \qquad [2.26]$$
$$n = 209,$$

where the intercept and slope estimates have been rounded to three decimal places; we use "*salary* hat" to indicate that this is an estimated equation. How do we interpret the equation? First, if the return on equity is zero, $roe = 0$, then the predicted *salary* is the intercept, 963.191, which equals \$963,191 because *salary* is measured in thousands. Next, we can write the predicted change in salary as a function of the change in *roe*: $\widehat{\Delta salary} = 18.501\,(\Delta roe)$. This means that if the return on equity increases by one percentage point, $\Delta roe = 1$, then *salary* is predicted to change by about 18.5, or \$18,500. Because (2.26) is a linear equation, this is the estimated change regardless of the initial salary.

We can easily use (2.26) to compare predicted salaries at different values of *roe*. Suppose $roe = 30$. Then $\widehat{salary} = 963.191 + 18.501(30) = 1{,}518{,}221$, which is just over \$1.5 million. However, this does *not* mean that a particular CEO whose firm had a $roe = 30$ earns \$1,518,221. Many other factors affect salary. This is just our prediction from the OLS regression line (2.26). The estimated line is graphed in Figure 2.5, along with the population regression function $\mathrm{E}(salary|roe)$. We will never know the PRF, so we cannot tell how close the SRF is to the PRF. Another sample of data will give a different regression line, which may or may not be closer to the population regression line.

FIGURE 2.5 The OLS regression line $\widehat{salary} = 963.191 + 18.501\ roe$ and the (unknown) population regression function.



EXAMPLE 2.4 Wage and Education

For the population of people in the workforce in 1976, let $y = wage$, where *wage* is measured in dollars per hour. Thus, for a particular person, if $wage = 6.75$, the hourly *wage* is \$6.75. Let $x = educ$ denote years of schooling; for example, $educ = 12$ corresponds to a complete high school education. Because the average wage in the sample is \$5.90, the Consumer Price Index indicates that this amount is equivalent to \$24.90 in 2016 dollars.

Using the data in WAGE1 where $n = 526$ individuals, we obtain the following OLS regression line (or sample regression function):

$$\widehat{wage} = -0.90 + 0.54\ educ \qquad [2.27]$$

$$n = 526.$$

We must interpret this equation with caution. The intercept of −0.90 literally means that a person with no education has a predicted hourly wage of −90¢ an hour. This, of course, is silly. It turns out that only 18 people in the sample of 526 have less than eight years of education. Consequently, it is not surprising that the regression line does poorly at very low levels of education. For a person with eight years of education, the predicted wage is $\widehat{wage} = -0.90 + 0.54(8) = 3.42$, or \$3.42 per hour (in 1976 dollars).

The slope estimate in (2.27) implies that one more year of education increases hourly wage by 54¢ an hour. Therefore, four more years of education increase the predicted wage by $4(0.54) = 2.16$, or \$2.16 per hour. These are fairly large effects.

GOING FURTHER 2.2

The estimated wage from (2.27), when *educ* = 8, is \$3.42 in 1976 dollars. What is this value in 2016 dollars? (*Hint*: You have enough information in Example 2.4 to answer this question.)

Because of the linear nature of (2.27), another year of education increases the wage by the same amount, regardless of the initial level of education. In Section 2-4, we discuss some methods that allow for nonconstant marginal effects of our explanatory variables.

EXAMPLE 2.5 **Voting Outcomes and Campaign Expenditures**

The file VOTE1 contains data on election outcomes and campaign expenditures for 173 two-party races for the U.S. House of Representatives in 1988. There are two candidates in each race, A and B. Let *voteA* be the percentage of the vote received by Candidate A and *shareA* be the percentage of total campaign expenditures accounted for by Candidate A. Many factors other than *shareA* affect the election outcome (including the quality of the candidates and possibly the dollar amounts spent by A and B). Nevertheless, we can estimate a simple regression model to find out whether spending more relative to one's challenger implies a higher percentage of the vote.

The estimated equation using the 173 observations is

$$\widehat{voteA} = 26.81 + 0.464\, shareA \qquad [2.28]$$

$$n = 173.$$

This means that if Candidate A's share of spending increases by one percentage point, Candidate A receives almost one-half a percentage point (0.464) more of the total vote. Whether or not this is a causal effect is unclear, but it is not unbelievable. If $shareA = 50$, *voteA* is predicted to be about 50, or half the vote.

GOING FURTHER 2.3

In Example 2.5, what is the predicted vote for Candidate A if *shareA* = 60 (which means 60%)? Does this answer seem reasonable?

In some cases, regression analysis is not used to determine causality but to simply look at whether two variables are positively or negatively related, much like a standard correlation analysis. An example of this occurs in Computer Exercise C3, where you are asked to use data from Biddle and Hamermesh (1990) on time spent sleeping and working to investigate the tradeoff between these two factors.

2-2a A Note on Terminology

In most cases, we will indicate the estimation of a relationship through OLS by writing an equation such as (2.26), (2.27), or (2.28). Sometimes, for the sake of brevity, it is useful to indicate that an OLS regression has been run without actually writing out the equation. We will often indicate that equation (2.23) has been obtained by OLS in saying that we *run the regression of*

$$y \text{ on } x, \qquad [2.29]$$

or simply that we *regress* y on x. The positions of y and x in (2.29) indicate which is the dependent variable and which is the independent variable: We always regress the dependent variable on the independent variable. For specific applications, we replace y and x with their names. Thus, to obtain (2.26), we regress *salary* on *roe*, or to obtain (2.28), we regress *voteA* on *shareA*.

When we use such terminology in (2.29), we will always mean that we plan to estimate the intercept, $\hat{\beta}_0$, along with the slope, $\hat{\beta}_1$. This case is appropriate for the vast majority of applications.

Occasionally, we may want to estimate the relationship between y and x *assuming* that the intercept is zero (so that $x = 0$ implies that $\hat{y} = 0$); we cover this case briefly in Section 2-6. Unless explicitly stated otherwise, we always estimate an intercept along with a slope.

2-3 Properties of OLS on Any Sample of Data

In the previous section, we went through the algebra of deriving the formulas for the OLS intercept and slope estimates. In this section, we cover some further algebraic properties of the fitted OLS regression line. The best way to think about these properties is to remember that they hold, by construction, for *any* sample of data. The harder task—considering the properties of OLS across all possible random samples of data—is postponed until Section 2-5.

Several of the algebraic properties we are going to derive will appear mundane. Nevertheless, having a grasp of these properties helps us to figure out what happens to the OLS estimates and related statistics when the data are manipulated in certain ways, such as when the measurement units of the dependent and independent variables change.

2-3a Fitted Values and Residuals

We assume that the intercept and slope estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, have been obtained for the given sample of data. Given $\hat{\beta}_0$ and $\hat{\beta}_1$, we can obtain the fitted value $\hat{y}_i$ for each observation. [This is given by equation (2.20).] By definition, each fitted value of $\hat{y}_i$ is on the OLS regression line. The OLS residual associated with observation i, $\hat{u}_i$, is the difference between y_i and its fitted value, as given in equation (2.21). If $\hat{u}_i$ is positive, the line underpredicts y_i; if $\hat{u}_i$ is negative, the line overpredicts y_i. The ideal case for observation i is when $\hat{u}_i = 0$, but in most cases, *every* residual is not equal to zero. In other words, none of the data points must actually lie on the OLS line.

EXAMPLE 2.6 **CEO Salary and Return on Equity**

Table 2.2 contains a listing of the first 15 observations in the CEO data set, along with the fitted values, called *salaryhat*, and the residuals, called *uhat*.

The first four CEOs have lower salaries than what we predicted from the OLS regression line (2.26); in other words, given only the firm's *roe*, these CEOs make less than what we predicted. As can be seen from the positive *uhat*, the fifth CEO makes more than predicted from the OLS regression line.

2-3b Algebraic Properties of OLS Statistics

There are several useful algebraic properties of OLS estimates and their associated statistics. We now cover the three most important of these.

(1) The sum, and therefore the sample average of the OLS residuals, is zero. Mathematically,

$$\sum_{i=1}^{n} \hat{u}_i = 0. \qquad [2.30]$$

This property needs no proof; it follows immediately from the OLS first order condition (2.14), when we remember that the residuals are defined by $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. In other words, the OLS estimates

TABLE 2.2 Fitted Values and Residuals for the First 15 CEOs

obsno	roe	salary	$\widehat{salary}$	$\hat{u}$
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	−616.7726
11	25.9	567	1442.372	−875.3721
12	26.8	933	1459.023	−526.0231
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	−438.7678
15	56.3	2011	2004.808	6.191895

$\hat{\beta}_0$ and $\hat{\beta}_1$ are *chosen* to make the residuals add up to zero (for any data set). This says nothing about the residual for any particular observation i.

(2) The sample covariance between the regressors and the OLS residuals is zero. This follows from the first order condition (2.15), which can be written in terms of the residuals as

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0. \qquad [2.31]$$

The sample average of the OLS residuals is zero, so the left-hand side of (2.31) is proportional to the sample covariance between x_i and $\hat{u}_i$.

(3) The point $(\bar{x}, \bar{y})$ is always on the OLS regression line. In other words, if we take equation (2.23) and plug in $\bar{x}$ for x, then the predicted value is $\bar{y}$. This is exactly what equation (2.16) showed us.

EXAMPLE 2.7 Wage and Education

For the data in WAGE1, the average hourly wage in the sample is 5.90, rounded to two decimal places, and the average education is 12.56. If we plug $educ = 12.56$ into the OLS regression line (2.27), we get $\widehat{wage} = -0.90 + 0.54(12.56) = 5.8824$, which equals 5.9 when rounded to the first decimal place. These figures do not exactly agree because we have rounded the average wage and education, as well as the intercept and slope estimates. If we did not initially round any of the values, we would get the answers to agree more closely, but to little useful effect.

Writing each y_i as its fitted value, plus its residual, provides another way to interpret an OLS regression. For each i, write

$$y_i = \hat{y}_i + \hat{u}_i. \qquad [2.32]$$

From property (1), the average of the residuals is zero; equivalently, the sample average of the fitted values, $\hat{y}_i$, is the same as the sample average of the y_i, or $\bar{\hat{y}} = \bar{y}$. Further, properties (1) and (2) can be used to show that the sample covariance between $\hat{y}_i$ and $\hat{u}_i$ is zero. Thus, we can view OLS as decomposing each y_i into two parts, a fitted value and a residual. The fitted values and residuals are uncorrelated in the sample.

Define the **total sum of squares (SST)**, the **explained sum of squares (SSE)**, and the **residual sum of squares (SSR)** (also known as the sum of squared residuals), as follows:

$$\text{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2. \qquad [2.33]$$

$$\text{SSE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \qquad [2.34]$$

$$\text{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2. \qquad [2.35]$$

SST is a measure of the total sample variation in the y_i; that is, it measures how spread out the y_i are in the sample. If we divide SST by $n - 1$, we obtain the sample variance of y, as discussed in Math Refresher C. Similarly, SSE measures the sample variation in the $\hat{y}_i$ (where we use the fact that $\bar{\hat{y}} = \bar{y}$), and SSR measures the sample variation in the $\hat{u}_i$. The total variation in y can always be expressed as the sum of the explained variation and the unexplained variation SSR. Thus,

$$\text{SST} = \text{SSE} + \text{SSR}. \qquad [2.36]$$

Proving (2.36) is not difficult, but it requires us to use all of the properties of the summation operator covered in Math Refresher A. Write

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n} [\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\
&= \sum_{i=1}^{n} \hat{u}_i^2 + 2\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&= \text{SSR} + 2\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) + \text{SSE}.
\end{aligned}$$

Now, (2.36) holds if we show that

$$\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) = 0. \qquad [2.37]$$

But we have already claimed that the sample covariance between the residuals and the fitted values is zero, and this covariance is just (2.37) divided by $n - 1$. Thus, we have established (2.36).

Some words of caution about SST, SSE, and SSR are in order. There is no uniform agreement on the names or abbreviations for the three quantities defined in equations (2.33), (2.34), and (2.35). The total sum of squares is called either SST or TSS, so there is little confusion here. Unfortunately, the explained sum of squares is sometimes called the "regression sum of squares." If this term is given its natural abbreviation, it can easily be confused with the term "residual sum of squares." Some regression packages refer to the explained sum of squares as the "model sum of squares."

To make matters even worse, the residual sum of squares is often called the "error sum of squares." This is especially unfortunate because, as we will see in Section 2-5, the errors and the residuals are different quantities. Thus, we will always call (2.35) the residual sum of squares or the sum of squared residuals. We prefer to use the abbreviation SSR to denote the sum of squared residuals, because it is more common in econometric packages.

2-3c Goodness-of-Fit

So far, we have no way of measuring how well the explanatory or independent variable, x, explains the dependent variable, y. It is often useful to compute a number that summarizes how well the OLS regression line fits the data. In the following discussion, be sure to remember that we assume that an intercept is estimated along with the slope.

Assuming that the total sum of squares, SST, is not equal to zero—which is true except in the very unlikely event that all the y_i equal the same value—we can divide (2.36) by SST to get $1 = \text{SSE/SST} + \text{SSR/SST}$. The ***R*-squared** of the regression, sometimes called the **coefficient of determination**, is defined as

$$R^2 \equiv \text{SSE/SST} = 1 - \text{SSR/SST}. \qquad [2.38]$$

R^2 is the ratio of the explained variation compared to the total variation; thus, it is interpreted as the *fraction of the sample variation in y that is explained by x*. The second equality in (2.38) provides another way for computing R^2.

From (2.36), the value of R^2 is always between zero and one, because SSE can be no greater than SST. When interpreting R^2, we usually multiply it by 100 to change it into a percent: $100 \cdot R^2$ is the *percentage of the sample variation in y that is explained by x*.

If the data points all lie on the same line, OLS provides a perfect fit to the data. In this case, $R^2 = 1$. A value of R^2 that is nearly equal to zero indicates a poor fit of the OLS line: very little of the variation in the y_i is captured by the variation in the $\hat{y}_i$ (which all lie on the OLS regression line). In fact, it can be shown that R^2 is equal to the *square* of the sample correlation coefficient between y_i and $\hat{y}_i$. This is where the term "*R*-squared" came from. (The letter R was traditionally used to denote an estimate of a population correlation coefficient, and its usage has survived in regression analysis.)

EXAMPLE 2.8 **CEO Salary and Return on Equity**

In the CEO salary regression, we obtain the following:

$$\widehat{salary} = 963.191 + 18.501\, roe \qquad [2.39]$$

$$n = 209, R^2 = 0.0132.$$

We have reproduced the OLS regression line and the number of observations for clarity. Using the *R*-squared (rounded to four decimal places) reported for this equation, we can see how much of the variation in salary is actually explained by the return on equity. The answer is: not much. The firm's return on equity explains only about 1.3% of the variation in salaries for this sample of 209 CEOs. That means that 98.7% of the salary variations for these CEOs is left unexplained! This lack of explanatory power may not be too surprising because many other characteristics of both the firm and the individual CEO should influence salary; these factors are necessarily included in the errors in a simple regression analysis.

In the social sciences, low *R*-squareds in regression equations are not uncommon, especially for cross-sectional analysis. We will discuss this issue more generally under multiple regression analysis, but it is worth emphasizing now that a seemingly low *R*-squared does not necessarily mean that an OLS regression equation is useless. It is still possible that (2.39) is a good estimate of the ceteris paribus relationship between *salary* and *roe*; whether or not this is true does *not* depend directly on the size of *R*-squared. Students who are first learning econometrics tend to put too much weight on the size of the *R*-squared in evaluating regression equations. For now, be aware that using *R*-squared as the main gauge of success for an econometric analysis can lead to trouble.

Sometimes, the explanatory variable explains a substantial part of the sample variation in the dependent variable.

EXAMPLE 2.9 **Voting Outcomes and Campaign Expenditures**

In the voting outcome equation in (2.28), $R^2 = 0.856$. Thus, the share of campaign expenditures explains over 85% of the variation in the election outcomes for this sample. This is a sizable portion.

2-4 Units of Measurement and Functional Form

Two important issues in applied economics are (1) understanding how changing the units of measurement of the dependent and/or independent variables affects OLS estimates and (2) knowing how to incorporate popular functional forms used in economics into regression analysis. The mathematics needed for a full understanding of functional form issues is reviewed in Math Refresher A.

2-4a The Effects of Changing Units of Measurement on OLS Statistics

In Example 2.3, we chose to measure annual salary in thousands of dollars, and the return on equity was measured as a percentage (rather than as a decimal). It is crucial to know how *salary* and *roe* are measured in this example in order to make sense of the estimates in equation (2.39).

We must also know that OLS estimates change in entirely expected ways when the units of measurement of the dependent and independent variables change. In Example 2.3, suppose that, rather than measuring salary in thousands of dollars, we measure it in dollars. Let *salardol* be salary in dollars (*salardol* = 845,761 would be interpreted as \$845,761). Of course, *salardol* has a simple relationship to the salary measured in thousands of dollars: $salardol = 1{,}000 \cdot salary$. We do not need to actually run the regression of *salardol* on *roe* to know that the estimated equation is:

$$\widehat{salardol} = 963{,}191 + 18{,}501\, roe. \qquad [2.40]$$

We obtain the intercept and slope in (2.40) simply by multiplying the intercept and the slope in (2.39) by 1,000. This gives equations (2.39) and (2.40) the *same* interpretation. Looking at (2.40), if $roe = 0$, then $\widehat{salardol} = 963{,}191$, so the predicted salary is \$963,191 [the same value we obtained from equation (2.39)]. Furthermore, if *roe* increases by one, then the predicted salary increases by \$18,501; again, this is what we concluded from our earlier analysis of equation (2.39).

Generally, it is easy to figure out what happens to the intercept and slope estimates when the dependent variable changes units of measurement. If the dependent variable is multiplied by the constant c—which means each value in the sample is multiplied by c—then the OLS intercept and slope estimates are also multiplied by c. (This assumes nothing has changed about the independent variable.) In the CEO salary example, $c = 1{,}000$ in moving from *salary* to *salardol*.

GOING FURTHER 2.4

Suppose that salary is measured in hundreds of dollars, rather than in thousands of dollars, say, *salarhun*. What will be the OLS intercept and slope estimates in the regression of *salarhun* on *roe*?

We can also use the CEO salary example to see what happens when we change the units of measurement of the independent variable. Define $roedec = roe/100$ to be the decimal equivalent of *roe*; thus, $roedec = 0.23$ means a return on equity of 23%. To focus on changing the units of measurement

of the independent variable, we return to our original dependent variable, *salary*, which is measured in thousands of dollars. When we regress *salary* on *roedec*, we obtain

$$\widehat{salary} = 963.191 + 1{,}850.1\, roedec. \qquad [2.41]$$

The coefficient on *roedec* is 100 times the coefficient on *roe* in (2.39). This is as it should be. Changing *roe* by one percentage point is equivalent to $\Delta roedec = 0.01$. From (2.41), if $\Delta roedec = 0.01$, then $\widehat{\Delta salary} = 1{,}850.1(0.01) = 18.501$, which is what is obtained by using (2.39). Note that, in moving from (2.39) to (2.41), the independent variable was divided by 100, and so the OLS slope estimate was multiplied by 100, preserving the interpretation of the equation. Generally, if the independent variable is divided or multiplied by some nonzero constant, c, then the OLS slope coefficient is multiplied or divided by c, respectively.

The intercept has not changed in (2.41) because $roedec = 0$ still corresponds to a zero return on equity. In general, changing the units of measurement of only the independent variable does not affect the intercept.

In the previous section, we defined R-squared as a goodness-of-fit measure for OLS regression. We can also ask what happens to R^2 when the unit of measurement of either the independent or the dependent variable changes. Without doing any algebra, we should know the result: the goodness-of-fit of the model should not depend on the units of measurement of our variables. For example, the amount of variation in salary explained by the return on equity should not depend on whether salary is measured in dollars or in thousands of dollars or on whether return on equity is a percentage or a decimal. This intuition can be verified mathematically: using the definition of R^2, it can be shown that R^2 is, in fact, invariant to changes in the units of y or x.

2-4b Incorporating Nonlinearities in Simple Regression

So far, we have focused on *linear* relationships between the dependent and independent variables. As we mentioned in Chapter 1, linear relationships are not nearly general enough for all economic applications. Fortunately, it is rather easy to incorporate many nonlinearities into simple regression analysis by appropriately defining the dependent and independent variables. Here, we will cover two possibilities that often appear in applied work.

In reading applied work in the social sciences, you will often encounter regression equations where the dependent variable appears in logarithmic form. Why is this done? Recall the wage-education example, where we regressed hourly wage on years of education. We obtained a slope estimate of 0.54 [see equation (2.27)], which means that each additional year of education is predicted to increase hourly wage by 54 cents. Because of the linear nature of (2.27), 54 cents is the increase for either the first year of education or the twentieth year; this may not be reasonable.

Probably a better characterization of how wage changes with education is that each year of education increases wage by a constant *percentage*. For example, an increase in education from 5 years to 6 years increases wage by, say, 8% (ceteris paribus), and an increase in education from 11 to 12 years also increases wage by 8%. A model that gives (approximately) a constant percentage effect is

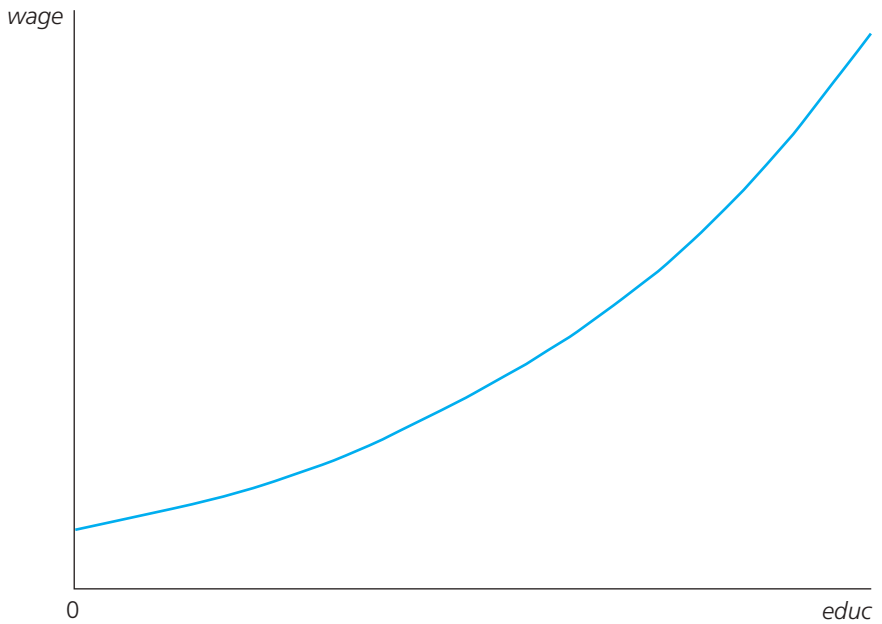
$$\log(wage) = \beta_0 + \beta_1 educ + u, \qquad [2.42]$$

where $\log(\cdot)$ denotes the *natural* logarithm. (See Math Refresher A for a review of logarithms.) In particular, if $\Delta u = 0$, then

$$\%\Delta wage \approx (100 \cdot \beta_1)\Delta educ. \qquad [2.43]$$

Notice how we multiply β_1 by 100 to get the percentage change in *wage* given one additional year of education. Because the percentage change in *wage* is the same for each additional year of education, the change in *wage* for an extra year of education *increases* as education increases; in other words, (2.42) implies an *increasing* return to education. By exponentiating (2.42), we can write $wage = \exp(\beta_0 + \beta_1 educ + u)$. This equation is graphed in Figure 2.6, with $u = 0$.

FIGURE 2.6 $wage = exp(\beta_0 + \beta_1 educ)$, with $\beta_1 > 0$.



EXAMPLE 2.10 **A Log Wage Equation**

Using the same data as in Example 2.4, but using log(*wage*) as the dependent variable, we obtain the following relationship:

$$\widehat{\log(wage)} = 0.584 + 0.083\, educ \qquad [2.44]$$

$$n = 526,\ R^2 = 0.186.$$

The coefficient on *educ* has a percentage interpretation when it is multiplied by 100: $\widehat{wage}$ increases by 8.3% for every additional year of education. This is what economists mean when they refer to the "return to another year of education."

It is important to remember that the main reason for using the log of *wage* in (2.42) is to impose a constant percentage effect of education on *wage*. Once equation (2.44) is obtained, the natural log of *wage* is rarely mentioned. In particular, it is *not* correct to say that another year of education increases log(*wage*) by 8.3%.

The intercept in (2.44) is not very meaningful, because it gives the predicted log(*wage*), when $educ = 0$. The *R*-squared shows that *educ* explains about 18.6% of the variation in log(*wage*) (*not* *wage*). Finally, equation (2.44) might not capture all of the nonlinearity in the relationship between wage and schooling. If there are "diploma effects," then the twelfth year of education—graduation from high school—could be worth much more than the eleventh year. We will learn how to allow for this kind of nonlinearity in Chapter 7.

Estimating a model such as (2.42) is straightforward when using simple regression. Just define the dependent variable, y, to be $y = \log(wage)$. The independent variable is represented by $x = educ$. The mechanics of OLS are the same as before: the intercept and slope estimates are given by the formulas (2.17) and (2.19). In other words, we obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ from the OLS regression of log(*wage*) on *educ*.

Another important use of the natural log is in obtaining a **constant elasticity model**.

EXAMPLE 2.11 **CEO Salary and Firm Sales**

We can estimate a constant elasticity model relating CEO salary to firm sales. The data set is the same one used in Example 2.3, except we now relate *salary* to *sales*. Let *sales* be annual firm sales, measured in millions of dollars. A constant elasticity model is

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u, \qquad [2.45]$$

where β_1 is the elasticity of *salary* with respect to *sales*. This model falls under the simple regression model by defining the dependent variable to be $y = \log(salary)$ and the independent variable to be $x = \log(sales)$. Estimating this equation by OLS gives

$$\widehat{\log(salary)} = 4.822 + 0.257 \log(sales) \qquad [2.46]$$

$$n = 209, R^2 = 0.211.$$

The coefficient of $\log(sales)$ is the estimated elasticity of *salary* with respect to *sales*. It implies that a 1% increase in firm sales increases CEO salary by about 0.257%—the usual interpretation of an elasticity.

The two functional forms covered in this section will often arise in the remainder of this text. We have covered models containing natural logarithms here because they appear so frequently in applied work. The interpretation of such models will not be much different in the multiple regression case.

It is also useful to note what happens to the intercept and slope estimates if we change the units of measurement of the dependent variable when it appears in logarithmic form. Because the change to logarithmic form approximates a proportionate change, it makes sense that *nothing* happens to the slope. We can see this by writing the rescaled variable as $c_1 y_i$ for each observation i. The original equation is $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$. If we add $\log(c_1)$ to both sides, we get $\log(c_1) + \log(y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$, *or* $\log(c_1 y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$. (Remember that the sum of the logs is equal to the log of their product, as shown in Math Refresher A.) Therefore, the slope is still β_1, but the intercept is now $\log(c_1) + \beta_0$. Similarly, if the independent variable is $\log(x)$, and we change the units of measurement of x before taking the log, the slope remains the same, but the intercept changes. You will be asked to verify these claims in Problem 9.

We end this subsection by summarizing four combinations of functional forms available from using either the original variable or its natural log. In Table 2.3, x and y stand for the variables in their original form. The model with y as the dependent variable and x as the independent variable is called the *level-level* model because each variable appears in its level form. The model with $\log(y)$ as the dependent variable and x as the independent variable is called the *log-level* model. We will not explicitly discuss the *level-log* model here, because it arises less often in practice. In any case, we will see examples of this model in later chapters.

The last column in Table 2.3 gives the interpretation of β_1. In the log-level model, $100 \cdot \beta_1$ is sometimes called the **semi-elasticity** of y with respect to x. As we mentioned in Example 2.11, in the log-log model, β_1 is the **elasticity** of y with respect to x. Table 2.3 warrants careful study, as we will refer to it often in the remainder of the text.

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

2-4c The Meaning of "Linear" Regression

The simple regression model that we have studied in this chapter is also called the simple *linear* regression model. Yet, as we have just seen, the general model also allows for certain *nonlinear* relationships. So what does "linear" mean here? You can see by looking at equation (2.1) that $y = \beta_0 + \beta_1 x + u$. The key is that this equation is linear in the *parameters* β_0 and β_1. There are no restrictions on how y and x relate to the original explained and explanatory variables of interest. As we saw in Examples 2.10 and 2.11, y and x can be natural logs of variables, and this is quite common in applications. But we need not stop there. For example, nothing prevents us from using simple regression to estimate a model such as $cons = \beta_0 + \beta_1 \sqrt{inc} + u$, where *cons* is annual consumption and *inc* is annual income.

Whereas the mechanics of simple regression do not depend on how y and x are defined, the interpretation of the coefficients does depend on their definitions. For successful empirical work, it is much more important to become proficient at interpreting coefficients than to become efficient at computing formulas such as (2.19). We will get much more practice with interpreting the estimates in OLS regression lines when we study multiple regression.

Plenty of models *cannot* be cast as a linear regression model because they are not linear in their parameters; an example is $cons = 1/(\beta_0 + \beta_1 inc) + u$. Estimation of such models takes us into the realm of the *nonlinear regression model*, which is beyond the scope of this text. For most applications, choosing a model that can be put into the linear regression framework is sufficient.

2-5 Expected Values and Variances of the OLS Estimators

In Section 2-1, we defined the population model $y = \beta_0 + \beta_1 x + u$, and we claimed that the key assumption for simple regression analysis to be useful is that the expected value of u given any value of x is zero. In Sections 2-2, 2-3, and 2-4, we discussed the algebraic properties of OLS estimation. We now return to the population model and study the *statistical* properties of OLS. In other words, we now view $\hat{\beta}_0$ and $\hat{\beta}_1$ as *estimators* for the parameters β_0 and β_1 that appear in the population model. This means that we will study properties of the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ over different random samples from the population. (Math Refresher C contains definitions of estimators and reviews some of their important properties.)

2-5a Unbiasedness of OLS

We begin by establishing the unbiasedness of OLS under a simple set of assumptions. For future reference, it is useful to number these assumptions using the prefix "SLR" for simple linear regression. The first assumption defines the population model.

Assumption SLR.1 Linear in Parameters

In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as

$$y = \beta_0 + \beta_1 x + u, \qquad [2.47]$$

where β_0 and β_1 are the population intercept and slope parameters, respectively.

To be realistic, y, x, and u are all viewed as random variables in stating the population model. We discussed the interpretation of this model at some length in Section 2-1 and gave several examples. In the previous section, we learned that equation (2.47) is not as restrictive as it initially seems; by choosing

y and x appropriately, we can obtain interesting nonlinear relationships (such as constant elasticity models).

We are interested in using data on y and x to estimate the parameters β_0 and, especially, β_1. We assume that our data were obtained as a random sample. (See Math Refresher C for a review of random sampling.)

Assumption SLR.2 **Random Sampling**

We have a random sample of size n, $\{(x_i, y_i): i = 1, 2, \ldots, n\}$, following the population model in equation (2.47).

We will have to address failure of the random sampling assumption in later chapters that deal with time series analysis and sample selection problems. Not all cross-sectional samples can be viewed as outcomes of random samples, but many can be.

We can write (2.47) in terms of the random sample as

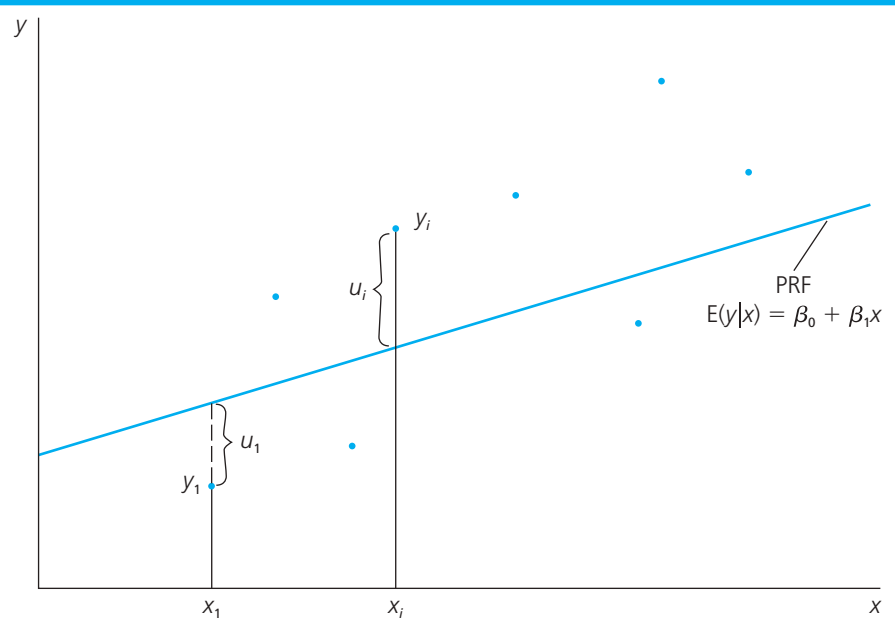
$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, \ldots, n, \qquad [2.48]$$

where u_i is the error or disturbance for observation i (for example, person i, firm i, city i, and so on). Thus, u_i contains the unobservables for observation i that affect y_i. The u_i should not be confused with the residuals, $\hat{u}_i$, that we defined in Section 2-3. Later on, we will explore the relationship between the errors and the residuals. For interpreting β_0 and β_1 in a particular application, (2.47) is most informative, but (2.48) is also needed for some of the statistical derivations.

The relationship (2.48) can be plotted for a particular outcome of data as shown in Figure 2.7.

As we already saw in Section 2-2, the OLS slope and intercept estimates are not defined unless we have some sample variation in the explanatory variable. We now add variation in the x_i to our list of assumptions.

FIGURE 2.7 Graph of $y_i = \beta_0 + \beta_1 x_i + u_i$.



Assumption SLR.3 **Sample Variation in the Explanatory Variable**

The sample outcomes on x, namely, $\{x_i, i = 1, \ldots, n\}$, are not all the same value.

This is a very weak assumption—certainly not worth emphasizing, but needed nevertheless. If x varies in the population, random samples on x will typically contain variation, unless the population variation is minimal or the sample size is small. Simple inspection of summary statistics on x_i reveals whether Assumption SLR.3 fails: if the sample standard deviation of x_i is zero, then Assumption SLR.3 fails; otherwise, it holds.

Finally, in order to obtain unbiased estimators of β_0 and β_1, we need to impose the zero conditional mean assumption that we discussed in some detail in Section 2-1. We now explicitly add it to our list of assumptions.

Assumption SLR.4 **Zero Conditional Mean**

The error u has an expected value of zero given any value of the explanatory variable. In other words,

$$\mathrm{E}(u|x) = 0.$$

For a random sample, this assumption implies that $\mathrm{E}(u_i|x_i) = 0$, for all $i = 1, 2, \ldots, n$.

In addition to restricting the relationship between u and x in the population, the zero conditional mean assumption—coupled with the random sampling assumption—allows for a convenient technical simplification. In particular, we can derive the statistical properties of the OLS estimators as *conditional* on the values of the x_i in our sample. Technically, in statistical derivations, conditioning on the sample values of the independent variable is the same as treating the x_i as *fixed in repeated samples*, which we think of as follows. We first choose n sample values for $x_1, x_2, \ldots, x_n$. (These can be repeated.) Given these values, we then obtain a sample on y (effectively by obtaining a random sample of the u_i). Next, another sample of y is obtained, using the *same* values for $x_1, x_2, \ldots, x_n$. Then another sample of y is obtained, again using the same $x_1, x_2, \ldots, x_n$. And so on.

The fixed-in-repeated-samples scenario is not very realistic in nonexperimental contexts. For instance, in sampling individuals for the wage-education example, it makes little sense to think of choosing the values of *educ* ahead of time and then sampling individuals with those particular levels of education. Random sampling, where individuals are chosen randomly and their wage and education are both recorded, is representative of how most data sets are obtained for empirical analysis in the social sciences. Once we *assume* that $\mathrm{E}(u|x) = 0$, and we have random sampling, nothing is lost in derivations by treating the x_i as nonrandom. The danger is that the fixed-in-repeated-samples assumption *always* implies that u_i and x_i are independent. In deciding when simple regression analysis is going to produce unbiased estimators, it is critical to think in terms of Assumption SLR.4.

Now, we are ready to show that the OLS estimators are unbiased. To this end, we use the fact that $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$ (see Math Refresher A) to write the OLS slope estimator in equation (2.19) as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad [2.49]$$

Because we are now interested in the behavior of $\hat{\beta}_1$ across all possible samples, $\hat{\beta}_1$ is properly viewed as a random variable.

We can write $\hat{\beta}_1$ in terms of the population coefficient and errors by substituting the right-hand side of (2.48) into (2.49). We have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\text{SST}_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\text{SST}_x}, \qquad [2.50]$$

where we have defined the total variation in x_i as $\text{SST}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ to simplify the notation. (This is not quite the sample variance of the x_i because we do not divide by $n - 1$.) Using the algebra of the summation operator, write the numerator of $\hat{\beta}_1$ as

$$\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 + \sum_{i=1}^{n}(x_i - \bar{x})\beta_1 x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i$$
$$= \beta_0\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i. \qquad [2.51]$$

As shown in Math Refresher A, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \text{SST}_x$. Therefore, we can write the numerator of $\hat{\beta}_1$ as $\beta_1\text{SST}_x + \sum_{i=1}^{n}(x_i - \bar{x})u_i$. Putting this over the denominator gives

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\text{SST}_x} = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n}d_i u_i, \qquad [2.52]$$

where $d_i = x_i - \bar{x}$. We now see that the estimator $\hat{\beta}_1$ equals the population slope, β_1, plus a term that is a linear combination in the errors $[u_1, u_2, \ldots, u_n]$. Conditional on the values of x_i, the randomness in $\hat{\beta}_1$ is due entirely to the errors in the sample. The fact that these errors are generally different from zero is what causes $\hat{\beta}_1$ to differ from β_1.

Using the representation in (2.52), we can prove the first important statistical property of OLS.

THEOREM 2.1

UNBIASEDNESS OF OLS:

Using Assumptions SLR.1 through SLR.4,

$$\text{E}(\hat{\beta}_0) = \beta_0 \text{ and } \text{E}(\hat{\beta}_1) = \beta_1, \qquad [2.53]$$

for any values of β_0 and β_1. In other words, $\hat{\beta}_0$ is unbiased for β_0, and $\hat{\beta}_1$ is unbiased for β_1.

PROOF: In this proof, the expected values are conditional on the sample values of the independent variable. Because SST_x and d_i are functions only of the x_i, they are nonrandom in the conditioning. Therefore, from (2.52), and keeping the conditioning on $\{x_1, x_2, \ldots, x_n\}$ implicit, we have

$$\text{E}(\hat{\beta}_1) = \beta_1 + \text{E}[(1/\text{SST}_x)\sum_{i=1}^{n}d_i u_i] = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n}\text{E}(d_i u_i)$$
$$= \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n}d_i\text{E}(U_i) = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n}d_i \cdot 0 = \beta_1,$$

where we have used the fact that the expected value of each u_i (conditional on $\{x_1, x_2, \ldots, x_n\}$) is zero under Assumptions SLR.2 and SLR.4. Because unbiasedness holds for any outcome on $\{x_1, x_2, \ldots, x_n\}$, unbiasedness also holds without conditioning on $\{x_1, x_2, \ldots, x_n\}$.

The proof for $\hat{\beta}_0$ is now straightforward. Average (2.48) across i to get $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{u}$, and plug this into the formula for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \beta_0 + \beta_1\bar{x} + \bar{u} - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}.$$

Then, conditional on the values of the x_i,

$$E(\hat{\beta}_0) = \beta_0 + E[(\beta_1 - \hat{\beta}_1)\bar{x}] + E(\bar{u}) = \beta_0 + E[(\beta_1 - \hat{\beta}_1)]\bar{x},$$

because $E(\bar{u}) = 0$ by Assumptions SLR.2 and SLR.4. But, we showed that $E(\hat{\beta}_1) = \beta_1$, which implies that $E[(\hat{\beta}_1 - \beta_1)] = 0$. Thus, $E(\hat{\beta}_0) = \beta_0$. Both of these arguments are valid for any values of β_0 and β_1, and so we have established unbiasedness.

Remember that unbiasedness is a feature of the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$, which says nothing about the estimate that we obtain for a given sample. We hope that, if the sample we obtain is somehow "typical," then our estimate should be "near" the population value. Unfortunately, it is always possible that we could obtain an unlucky sample that would give us a point estimate far from β_1, and we can *never* know for sure whether this is the case. You may want to review the material on unbiased estimators in Math Refresher C, especially the simulation exercise in Table C.1 that illustrates the concept of unbiasedness.

Unbiasedness generally fails if any of our four assumptions fail. This means that it is important to think about the veracity of each assumption for a particular application. Assumption SLR.1 requires that y and x be linearly related, with an additive disturbance. This can certainly fail. But we also know that y and x can be chosen to yield interesting nonlinear relationships. Dealing with the failure of (2.47) requires more advanced methods that are beyond the scope of this text.

Later, we will have to relax Assumption SLR.2, the random sampling assumption, for time series analysis. But what about using it for cross-sectional analysis? Random sampling can fail in a cross section when samples are not representative of the underlying population; in fact, some data sets are constructed by intentionally oversampling different parts of the population. We will discuss problems of nonrandom sampling in Chapters 9 and 17.

As we have already discussed, Assumption SLR.3 almost always holds in interesting regression applications. Without it, we cannot even obtain the OLS estimates.

The assumption we should concentrate on for now is SLR.4. If SLR.4 holds, the OLS estimators are unbiased. Likewise, if SLR.4 fails, the OLS estimators generally will be *biased*. There are ways to determine the likely direction and size of the bias, which we will study in Chapter 3.

The possibility that x is correlated with u is almost always a concern in simple regression analysis with nonexperimental data, as we indicated with several examples in Section 2-1. Using simple regression when u contains factors affecting y that are also correlated with x can result in *spurious correlation*: that is, we find a relationship between y and x that is really due to other unobserved factors that affect y and also happen to be correlated with x.

EXAMPLE 2.12 Student Math Performance and the School Lunch Program

Let *math10* denote the percentage of tenth graders at a high school receiving a passing score on a standardized mathematics exam. Suppose we wish to estimate the effect of the federally funded school lunch program on student performance. If anything, we expect the lunch program to have a positive ceteris paribus effect on performance: all other factors being equal, if a student who is too poor to eat regular meals becomes eligible for the school lunch program, his or her performance should improve. Let *lnchprg* denote the percentage of students who are eligible for the lunch program. Then, a simple regression model is

$$math10 = \beta_0 + \beta_1 lnchprg + u, \qquad [2.54]$$

where u contains school and student characteristics that affect overall school performance. Using the data in MEAP93 on 408 Michigan high schools for the 1992–1993 school year, we obtain

$$\widehat{math10} = 32.14 - 0.319\, lnchprg$$
$$n = 408, R^2 = 0.171.$$

This equation predicts that if student eligibility in the lunch program increases by 10 percentage points, the percentage of students passing the math exam *falls* by about 3.2 percentage points. Do we really believe that higher participation in the lunch program actually *causes* worse performance? Almost certainly not. A better explanation is that the error term u in equation (2.54) is correlated with *lnchprg*. In fact, u contains factors such as the poverty rate of children attending school, which affects student performance and is highly correlated with eligibility in the lunch program. Variables such as school quality and resources are also contained in u, and these are likely correlated with *lnchprg*. It is important to remember that the estimate –0.319 is only for this particular sample, but its sign and magnitude make us suspect that u and x are correlated, so that simple regression is biased.

In addition to omitted variables, there are other reasons for x to be correlated with u in the simple regression model. Because the same issues arise in multiple regression analysis, we will postpone a systematic treatment of the problem until then.

2-5b Variances of the OLS Estimators

In addition to knowing that the sampling distribution of $\hat{\beta}_1$ is centered about β_1 ($\hat{\beta}_1$ is unbiased), it is important to know how far we can expect $\hat{\beta}_1$ to be away from β_1 on average. Among other things, this allows us to choose the best estimator among all, or at least a broad class of, unbiased estimators. The measure of spread in the distribution of $\hat{\beta}_1$ (and $\hat{\beta}_0$) that is easiest to work with is the variance or its square root, the standard deviation. (See Math Refresher C for a more detailed discussion.)

It turns out that the variance of the OLS estimators can be computed under Assumptions SLR.1 through SLR.4. However, these expressions would be somewhat complicated. Instead, we add an assumption that is traditional for cross-sectional analysis. This assumption states that the variance of the unobservable, u, conditional on x, is constant. This is known as the **homoskedasticity** or "constant variance" assumption.

Assumption SLR.5 Homoskedasticity

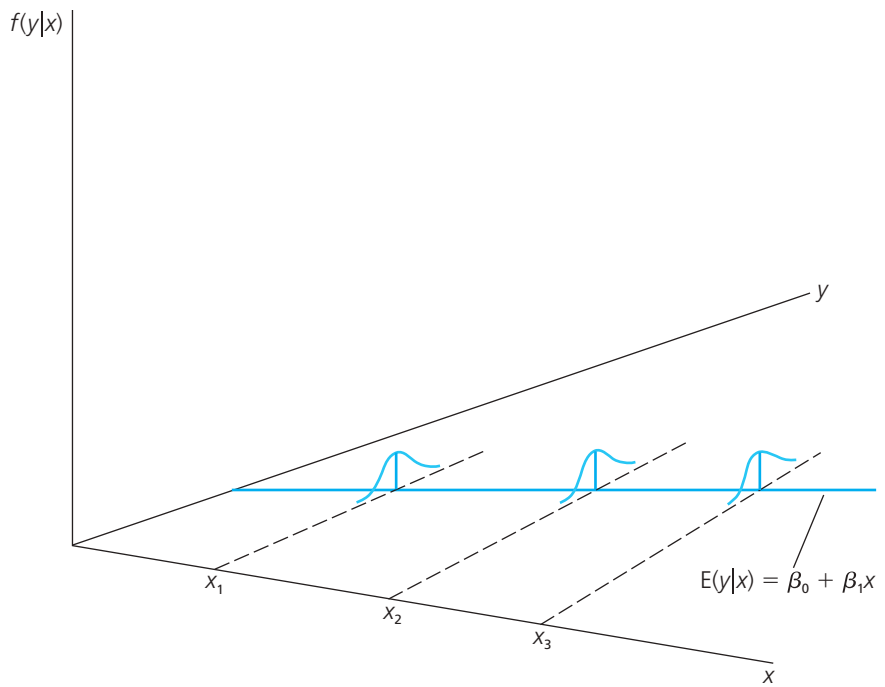
The error u has the same variance given any value of the explanatory variable. In other words,

$$\text{Var}(u|x) = \sigma^2.$$

We must emphasize that the homoskedasticity assumption is quite distinct from the zero conditional mean assumption, $\text{E}(u|x) = 0$. Assumption SLR.4 involves the *expected value* of u, while Assumption SLR.5 concerns the *variance* of u (both conditional on x). Recall that we established the unbiasedness of OLS without Assumption SLR.5: the homoskedasticity assumption plays *no* role in showing that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. We add Assumption SLR.5 because it simplifies the variance calculations for $\hat{\beta}_0$ and $\hat{\beta}_1$ and because it implies that ordinary least squares has certain efficiency properties, which we will see in Chapter 3. If we were to assume that u and x are *independent*, then the distribution of u given x does not depend on x, and so $\text{E}(u|x) = \text{E}(u) = 0$ and $\text{Var}(u|x) = \sigma^2$. But independence is sometimes too strong of an assumption.

Because $\text{Var}(u|x) = \text{E}(u^2|x) - [\text{E}(u|x)]^2$ and $\text{E}(u|x) = 0$, $\sigma^2 = \text{E}(u^2|x)$, which means σ^2 is also the *unconditional* expectation of u^2. Therefore, $\sigma^2 = \text{E}(u^2) = \text{Var}(u)$, because $\text{E}(u) = 0$. In other words, σ^2 is the *unconditional* variance of u, and so σ^2 is often called the **error variance** or disturbance variance. The square root of σ^2, σ, is the standard deviation of the error. A larger σ means that the distribution of the unobservables affecting y is more spread out.

FIGURE 2.8 The simple regression model under homoskedasticity.



It is often useful to write Assumptions SLR.4 and SLR.5 in terms of the conditional mean and conditional variance of y:

$$\mathrm{E}(y|x) = \beta_0 + \beta_1 x. \qquad [2.55]$$

$$\mathrm{Var}(y|x) = \sigma^2. \qquad [2.56]$$

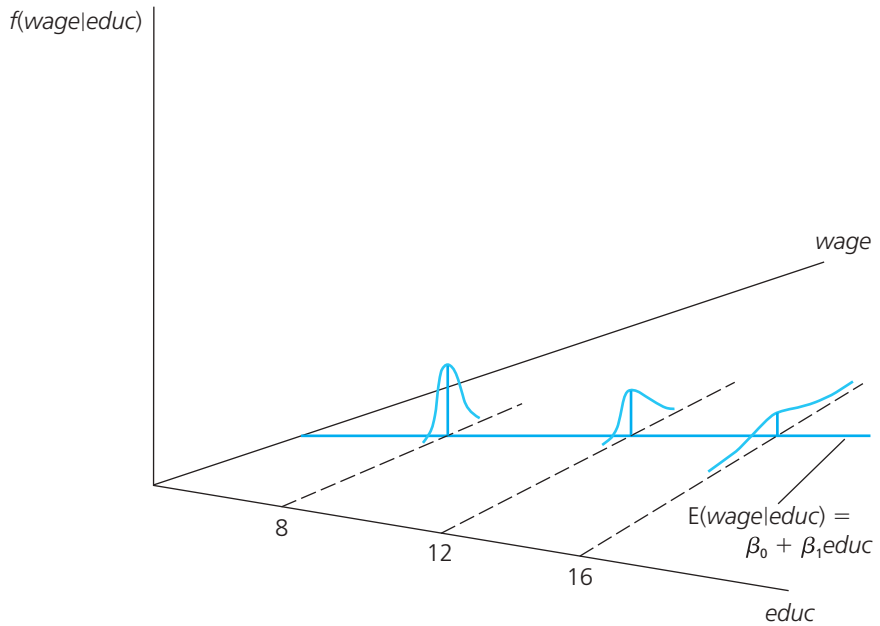
In other words, the conditional expectation of y given x is linear in x, but the variance of y given x is constant. This situation is graphed in Figure 2.8 where $\beta_0 > 0$ and $\beta_1 > 0$.

When $\mathrm{Var}(u|x)$ depends on x, the error term is said to exhibit **heteroskedasticity** (or nonconstant variance). Because $\mathrm{Var}(u|x) = \mathrm{Var}(y|x)$, heteroskedasticity is present whenever $\mathrm{Var}(y|x)$ is a function of x.

EXAMPLE 2.13 Heteroskedasticity in a Wage Equation

In order to get an unbiased estimator of the ceteris paribus effect of *educ* on *wage,* we must assume that $\mathrm{E}(u|educ) = 0$, and this implies $\mathrm{E}(wage|educ) = \beta_0 + \beta_1 educ$. If we also make the homoskedasticity assumption, then $\mathrm{Var}(u|educ) = \sigma^2$ does not depend on the level of education, which is the same as assuming $\mathrm{Var}(wage|educ) = \sigma^2$. Thus, while average wage is allowed to increase with education level—it is this rate of increase that we are interested in estimating—the *variability* in wage about its mean is assumed to be constant across all education levels. This may not be realistic. It is likely that people with more education have a wider variety of interests and job opportunities, which could lead to more wage variability at higher levels of education. People with very low levels of education have fewer opportunities and often must work at the minimum wage; this serves to reduce wage variability at low education levels. This situation is shown in Figure 2.9. Ultimately, whether Assumption SLR.5 holds is an empirical issue, and in Chapter 8 we will show how to test Assumption SLR.5.

FIGURE 2.9 Var(*wage*|*educ*) increasing with *educ*.



With the homoskedasticity assumption in place, we are ready to prove the following:

THEOREM 2.2

SAMPLING VARIANCES OF THE OLS ESTIMATORS

Under Assumptions SLR.1 through SLR.5,

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2/\text{SST}_x, \qquad [2.57]$$

and

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad [2.58]$$

where these are conditional on the sample values $\{x_1, \ldots, x_n\}$.

PROOF: We derive the formula for $\text{Var}(\hat{\beta}_1)$, leaving the other derivation as Problem 10. The starting point is equation (2.52): $\hat{\beta}_1 = \beta_1 + (1/\text{SST}_x)\sum_{i=1}^{n} d_i u_i$. Because β_1 is just a constant, and we are conditioning on the x_i, SST_x and $d_i = x_i - \bar{x}$ are also nonrandom. Furthermore, because the u_i are independent

random variables across i (by random sampling), the variance of the sum is the sum of the variances. Using these facts, we have

$$\text{Var}(\hat{\beta}_1) = (1/\text{SST}_x)^2\text{Var}\left(\sum_{i=1}^{n} d_i u_i\right) = (1/\text{SST}_x)^2\left(\sum_{i=1}^{n} d_i^2\text{Var}(u_i)\right)$$

$$= (1/\text{SST}_x)^2\left(\sum_{i=1}^{n} d_i^2\sigma^2\right) \quad [\text{because Var}(u_i) = \sigma^2 \text{ for all } i]$$

$$= \sigma^2(1/\text{SST}_x)^2\left(\sum_{i=1}^{n} d_i^2\right) = \sigma^2(1/\text{SST}_x)^2\text{SST}_x = \sigma^2/\text{SST}_x,$$

which is what we wanted to show.

Equations (2.57) and (2.58) are the "standard" formulas for simple regression analysis, which are invalid in the presence of heteroskedasticity. This will be important when we turn to confidence intervals and hypothesis testing in multiple regression analysis.

For most purposes, we are interested in $\text{Var}(\hat{\beta}_1)$. It is easy to summarize how this variance depends on the error variance, σ^2, and the total variation in $\{x_1, x_2, \ldots, x_n\}$, SST_x. First, the larger the error variance, the larger is $\text{Var}(\hat{\beta}_1)$. This makes sense because more variation in the unobservables affecting y makes it more difficult to precisely estimate β_1. On the other hand, more variability in the independent variable is preferred: as the variability in the x_i increases, the variance of $\hat{\beta}_1$ decreases. This also makes intuitive sense because the more spread out is the sample of independent variables, the easier it is to trace out the relationship between $\text{E}(y|x)$ and x; that is, the easier it is to estimate β_1. If there is little variation in the x_i, then it can be hard to pinpoint how $\text{E}(y|x)$ varies with x. As the sample size increases, so does the total variation in the x_i. Therefore, a larger sample size results in a smaller variance for $\hat{\beta}_1$.

GOING FURTHER 2.5

Show that, when estimating β_0, it is best to have $\bar{x} = 0$. What is $\text{Var}(\hat{\beta}_0)$ in this case? [*Hint:* For any sample of numbers, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2$, with equality only if $\bar{x} = 0$.]

This analysis shows that, if we are interested in β_1 and we have a choice, then we should choose the x_i to be as spread out as possible. This is sometimes possible with experimental data, but rarely do we have this luxury in the social sciences: usually, we must take the x_i that we obtain via random sampling. Sometimes, we have an opportunity to obtain larger sample sizes, although this can be costly.

For the purposes of constructing confidence intervals and deriving test statistics, we will need to work with the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_0$, $\text{sd}(\hat{\beta}_1)$ and $\text{sd}(\hat{\beta}_0)$. Recall that these are obtained by taking the square roots of the variances in (2.57) and (2.58). In particular, $\text{sd}(\hat{\beta}_1) = \sigma/\sqrt{\text{SST}_x}$, where σ is the square root of σ^2, and $\sqrt{\text{SST}_x}$ is the square root of SST_x.

2-5c Estimating the Error Variance

The formulas in (2.57) and (2.58) allow us to isolate the factors that contribute to $\text{Var}(\hat{\beta}_1)$ and $\text{Var}(\hat{\beta}_0)$. But these formulas are unknown, except in the extremely rare case that σ^2 is known. Nevertheless, we can use the data to estimate σ^2, which then allows us to estimate $\text{Var}(\hat{\beta}_1)$ and $\text{Var}(\hat{\beta}_0)$.

This is a good place to emphasize the difference between the *errors* (or disturbances) and the *residuals*, as this distinction is crucial for constructing an estimator of σ^2. Equation (2.48) shows how to write the population model in terms of a randomly sampled observation as $y_i = \beta_0 + \beta_1 x_i + u_i$, where u_i is the error for observation i. We can also express y_i in terms of its fitted value and residual as in equation (2.32): $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$. Comparing these two equations, we see that the error shows

up in the equation containing the *population* parameters, β_0 and β_1. On the other hand, the residuals show up in the *estimated* equation with $\hat{\beta}_0$ and $\hat{\beta}_1$. The errors are never observed, while the residuals are computed from the data.

We can use equations (2.32) and (2.48) to write the residuals as a function of the errors:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

or

$$\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)x_i. \qquad [2.59]$$

Although the expected value of $\hat{\beta}_0$ equals β_0, and similarly for $\hat{\beta}_1$, $\hat{u}_i$ is not the same as u_i. The difference between them does have an *expected value* of zero.

Now that we understand the difference between the errors and the residuals, we can return to estimating σ^2. First, $\sigma^2 = \mathrm{E}(u^2)$, so an unbiased "estimator" of σ^2 is $n^{-1}\sum_{i=1}^{n} u_i^2$. Unfortunately, this is not a true estimator, because we do not observe the errors u_i. But, we do have estimates of the u_i, namely, the OLS residuals $\hat{u}_i$. If we replace the errors with the OLS residuals, we have $n^{-1}\sum_{i=1}^{n} \hat{u}_i^2 = \mathrm{SSR}/n$. This *is* a true estimator, because it gives a computable rule for any sample of data on x and y. One slight drawback to this estimator is that it turns out to be biased (although for large n the bias is small). Because it is easy to compute an unbiased estimator, we use that instead.

The estimator SSR/n is biased essentially because it does not account for two restrictions that must be satisfied by the OLS residuals. These restrictions are given by the two OLS first order conditions:

$$\sum_{i=1}^{n} \hat{u}_i = 0, \quad \sum_{i=1}^{n} x_i \hat{u}_i = 0. \qquad [2.60]$$

One way to view these restrictions is this: if we know $n - 2$ of the residuals, we can always get the other two residuals by using the restrictions implied by the first order conditions in (2.60). Thus, there are only $n - 2$ **degrees of freedom** in the OLS residuals, as opposed to n degrees of freedom in the errors. It is important to understand that if we replace $\hat{u}_i$ with u_i in (2.60), the restrictions would no longer hold.

The unbiased estimator of σ^2 that we will use makes a degrees of freedom adjustment:

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} \hat{u}_i^2 = \mathrm{SSR}/(n-2). \qquad [2.61]$$

(This estimator is sometimes denoted as S^2, but we continue to use the convention of putting "hats" over estimators.)

THEOREM 2.3

UNBIASED ESTIMATION OF σ^2

Under Assumptions SLR.1 through SLR.5,

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2.$$

PROOF: If we average equation (2.59) across all i and use the fact that the OLS residuals average out to zero, we have $0 = \bar{u} - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)\bar{x}$; subtracting this from (2.59) gives $\hat{u}_i = (u_i - \bar{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. Therefore, $\hat{u}_i^2 = (u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2(x_i - \bar{x})^2 - 2(u_i - \bar{u})(\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. Summing across all i gives $\sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}(u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n} u_i(x_i - \bar{x})$. Now, the expected value of the first term is $(n - 1)\sigma^2$, something that is shown in Math Refresher C. The expected value of the second term is simply σ^2 because $\mathrm{E}[(\hat{\beta}_1 - \beta_1)^2] = \mathrm{Var}(\hat{\beta}_1) = \sigma^2/SST_x$. Finally, the third term can be written as $-2(\hat{\beta}_1 - \beta_1)^2 SST_x$; taking expectations gives $-2\sigma^2$. Putting these three terms together gives $\mathrm{E}(\sum_{i=1}^{n}\hat{u}_i^2) = (n - 1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n - 2)\sigma^2$, so that $\mathrm{E}[\mathrm{SSR}/(n - 2)] = \sigma^2$.

If $\hat{\sigma}^2$ is plugged into the variance formulas (2.57) and (2.58), then we have unbiased estimators of $\text{Var}(\hat{\beta}_1)$ and $\text{Var}(\hat{\beta}_0)$. Later on, we will need estimators of the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_0$, and this requires estimating σ. The natural estimator of σ is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \qquad [2.62]$$

and is called the **standard error of the regression (SER)**. (Other names for $\hat{\sigma}$ are the *standard error of the estimate* and the *root mean squared error*, but we will not use these.) Although $\hat{\sigma}$ is not an unbiased estimator of σ, we can show that it is a *consistent* estimator of σ (see Math Refresher C), and it will serve our purposes well.

The estimate $\hat{\sigma}$ is interesting because it is an estimate of the standard deviation in the unobservables affecting y; equivalently, it estimates the standard deviation in y after the effect of x has been taken out. Most regression packages report the value of $\hat{\sigma}$ along with the R-squared, intercept, slope, and other OLS statistics (under one of the several names listed above). For now, our primary interest is in using $\hat{\sigma}$ to estimate the standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$. Because $\text{sd}(\hat{\beta}_1) = \sigma/\sqrt{\text{SST}_x}$, the natural estimator of $\text{sd}(\hat{\beta}_1)$ is

$$\text{se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{SST}_x} = \hat{\sigma}/\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2};$$

this is called the **standard error of $\hat{\beta}_1$**. Note that $\text{se}(\hat{\beta}_1)$ is viewed as a random variable when we think of running OLS over different samples of y; this is true because $\hat{\sigma}$ varies with different samples. For a given sample, $\text{se}(\hat{\beta}_1)$ is a number, just as $\hat{\beta}_1$ is simply a number when we compute it from the given data.

Similarly, $\text{se}(\hat{\beta}_0)$ is obtained from $\text{sd}(\hat{\beta}_0)$ by replacing σ with $\hat{\sigma}$. The standard error of any estimate gives us an idea of how precise the estimator is. Standard errors play a central role throughout this text; we will use them to construct test statistics and confidence intervals for every econometric procedure we cover, starting in Chapter 4.

2-6 Regression through the Origin and Regression on a Constant

In rare cases, we wish to impose the restriction that, when $x = 0$, the expected value of y is zero. There are certain relationships for which this is reasonable. For example, if income (x) is zero, then income tax revenues (y) must also be zero. In addition, there are settings where a model that originally has a nonzero intercept is transformed into a model without an intercept.

Formally, we now choose a slope estimator, which we call $\tilde{\beta}_1$, and a line of the form

$$\tilde{y} = \tilde{\beta}_1 x, \qquad [2.63]$$

where the tildes over $\tilde{\beta}_1$ and $\tilde{y}$ are used to distinguish this problem from the much more common problem of estimating an intercept along with a slope. Obtaining (2.63) is called **regression through the origin** because the line (2.63) passes through the point $x = 0, \tilde{y} = 0$. To obtain the slope estimate in (2.63), we still rely on the method of ordinary least squares, which in this case minimizes the sum of squared residuals:

$$\sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x_i)^2. \qquad [2.64]$$

Using one-variable calculus, it can be shown that $\tilde{\beta}_1$ must solve the first order condition:

$$\sum_{i=1}^{n} x_i(y_i - \tilde{\beta}_1 x_i) = 0. \qquad [2.65]$$

From this, we can solve for $\tilde{\beta}_1$:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \qquad [2.66]$$

provided that not all the x_i are zero, a case we rule out.

Note how $\tilde{\beta}_1$ compares with the slope estimate when we also estimate the intercept (rather than set it equal to zero). These two estimates are the same if, and only if, $\bar{x} = 0$. [See equation (2.49) for $\hat{\beta}_1$.] Obtaining an estimate of β_1 using regression through the origin is not done very often in applied work, and for good reason: if the intercept $\beta_0 \neq 0$, then $\tilde{\beta}_1$ is a biased estimator of β_1. You will be asked to prove this in Problem 8.

In cases where regression through the origin is deemed appropriate, one must be careful in interpreting the *R*-squared that is typically reported with such regressions. Usually, unless stated otherwise, the *R*-squared is obtained without removing the sample average of $\{y_i: i = 1, \ldots, n\}$ in obtaining SST. In other words, the *R*-squared is computed as

$$1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x_i)^2}{\sum_{i=1}^{n} y_i^2}. \qquad [2.67]$$

The numerator here makes sense because it is the sum of squared residuals, but the denominator acts as if we know the average value of y in the population is zero. One reason this version of the *R*-squared is used is that if we use the usual total sum of squares, that is, we compute *R*-squared as

$$1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad [2.68]$$

it can actually be negative. If expression (2.68) is negative then it means that using the sample average $\bar{y}$ to predict y_i provides a better fit than using x_i in a regression through the origin. Therefore, (2.68) is actually more attractive than equation (2.67) because equation (2.68) tells us whether using x is better than ignoring x altogether.

This discussion about regression through the origin, and different ways to measure goodness-of-fit, prompts another question: what happens if we only regress on a constant? That is, what if we set the slope to zero (which means we need not even have an x) and estimate an intercept only? The answer is simple: the intercept is $\bar{y}$. This fact is usually shown in basic statistics, where it is shown that the constant that produces the smallest sum of squared deviations is always the sample average. In this light, equation (2.68) can be seen as comparing regression on x through the origin with regression only on a constant.

2-7 Regression on a Binary Explanatory Variable

Our discussion so far has centered on the case where the explanatory variable, x, has quantitative meanings. A few examples include years of schooling, return on equity for a firm, and the percentage of students at a school eligible for the federal free lunch program. We know how to interpret the slope coefficient in each case. We also discussed interpretation of the slope coefficient when we use the logarithmic transformations of the explained variable, the explanatory variable, or both.

Simple regression can also be applied to the case where x is a **binary variable**, often called a **dummy variable** in the context of regression analysis. As the name "binary variable" suggests, x takes on only two values, zero and one. These two values are used to put each unit in the population into one of two groups represented by $x = 0$ and $x = 1$. For example, we can use a binary variable to describe whether a worker participates in a job training program. In the spirit of giving our variables descriptive names, we might use *train* to indicate participation: $train = 1$ means a person participates; $train = 0$ means the person does not. Given a data set, we add an i subscript, as usual, so $train_i$ indicates job training status for a randomly drawn person i.

If we have a dependent or response variable, y, what does it mean to have a simple regression equation when x is binary? Consider again the equation

$$y = \beta_0 + \beta_1 x + u$$

but where now x is a binary variable. If we impose the zero conditional mean assumption SLR.4 then we obtain

$$\mathrm{E}(y|x) = \beta_0 + \beta_1 x + \mathrm{E}(u|x) = \beta_0 + \beta_1 x, \qquad [2.69]$$

just as in equation (2.8). The only difference now is that x can take on only two values. By plugging the values zero and one into (2.69), it is easily seen that

$$\mathrm{E}(y|x = 0) = \beta_0 \qquad [2.70]$$

$$\mathrm{E}(y|x = 1) = \beta_0 + \beta_1. \qquad [2.71]$$

It follows immediately that

$$\beta_1 = \mathrm{E}(y|x = 1) - \mathrm{E}(y|x = 0). \qquad [2.72]$$

In other words, β_1 is the difference in the average value of y over the subpopulations with $x = 1$ and $x = 0$. As with all simple regression analyses, this difference can be descriptive or, in a case discussed in the next subsection, β_1 can be a causal effect of an intervention or a program.

As an example, suppose that every worker in an hourly wage industry is put into one of two racial categories: white (or Caucasian) and nonwhite. (Clearly this is a very crude way to categorize race, but it has been used in some contexts.) Define the variable $white = 1$ if a person is classified as Caucasian and zero otherwise. Let *wage* denote hourly wage. Then

$$\beta_1 = \mathrm{E}(wage|white = 1) - \mathrm{E}(wage|white = 0)$$

is the difference in average hourly wages between white and nonwhite workers. Equivalently,

$$\mathrm{E}(wage|white) = \beta_0 + \beta_1 white.$$

Notice that β_1 always has the interpretation that it is the difference in average wages between whites and nonwhites. However, it does not necessarily measure wage discrimination because there are many legitimate reasons wages can differ, and some of those—such as education levels—could differ, on average, by race.

The mechanics of OLS do not change just because x is binary. Let $\{(x_i, y_i): i = 1, \ldots, n\}$ be the sample of size n. The OLS intercept and slope estimators are always given by (2.16) and (2.19), respectively. The residuals always have zero mean and are uncorrelated with the x_i in the sample. The definition of R-squared is unchanged. And so on. Nevertheless, because x_i is binary, the OLS estimates have a simple, sensible interpretation. Let $\bar{y}_0$ be the average of the y_i with $x_i = 0$ and $\bar{y}_1$ the average when $x_i = 1$. Problem 2.13 asks you to show that

$$\hat{\beta}_0 = \bar{y}_0 \qquad [2.73]$$

$$\hat{\beta}_0 = \bar{y}_1 - \bar{y}_0. \qquad [2.74]$$

For example, in the wage/race example, if we run the regression

$$wage_i \text{ on } white_i,\ i = 1, \ldots, n$$

then $\hat{\beta}_0 = \overline{wage}_0$, the average hourly wage for nonwhites, and $\hat{\beta}_1 = \overline{wage}_1 - \overline{wage}_0$, the difference in average hourly wages between whites and nonwhites. Generally, equation (2.74) shows that the "slope" in the regression is the difference in means, which is a standard estimator from basic statistics when comparing two groups.

The statistical properties of OLS are also unchanged when x is binary. In fact, nowhere is this ruled out in the statements of the assumptions. Assumption SLR.3 is satisfied provided we see some

zeros and some ones for x_i in our sample. For example, in the wage/race example, we need to observe some whites and some nonwhites in order to obtain $\hat{\beta}_1$.

As with any simple regression analysis, the main concern is the zero conditional mean assumption, SLR.4. In many cases, this condition will fail because x is systematically related to other factors that affect y, and those other factors are necessarily part of u. We alluded to this above in discussing differences in average hourly wage by race: education and workforce experience are two variables that affect hourly wage that could systematically differ by race. As another example, suppose we have data on SAT scores for students who did and did not take at least one SAT preparation course. Then x is a binary variable, say, *course,* and the outcome variable is the SAT score, *sat*. The decision to take the preparation course could be systematically related to other factors that are predictive of SAT scores, such as family income and parents' education. A comparison of average SAT scores between the two groups is unlikely to uncover the causal effect of the preparation course. The framework covered in the next subsection allows us to determine the special circumstances under which simple regression can uncover a causal effect.

2-7a Counterfactual Outcomes, Causality, and Policy Analysis

Having introduced the notion of a binary explanatory variable, now is a good time to provide a formal framework for studying counterfactual or potential outcomes, as touched on briefly in Chapter 1. We are particularly interested in defining a **causal effect** or **treatment effect**.

In the simplest case, we are interested in evaluating an intervention or policy that has only two states of the world: a unit is subjected to the intervention or not. In other words, those not subject to the intervention or new policy act as a **control group** and those subject to the intervention as the **treatment group**. Using the potential outcomes framework introduced in Chapter 1, for each unit i in the population we assume there are outcomes in both states of the world, $y_i(0)$ and $y_i(1)$. We will never observe any unit in both states of the world but we imagine each unit in both states. For example, in studying a job training program, a person does or does not participate. Then $y_i(0)$ is earnings if person i does not participate and $y_i(1)$ is labor earnings if i does participate. These outcomes are well defined before the program is even implemented.

The causal effect, somewhat more commonly called the treatment effect, of the intervention for unit i is simply

$$te_i = y_i(1) - y_i(0), \tag{2.75}$$

the difference between the two potential outcomes. There are a couple of noteworthy items about te_i. First, it is not observed for any unit i because it depends on both counterfactuals. Second, it can be negative, zero, or positive. It could be that the causal effect is negative for some units and positive for others.

We cannot hope to estimate te_i for each unit i. Instead, the focus is typically on the **average treatment effect (ATE)**, also called the **average causal effect (ACE)**. The ATE is simply the average of the treatment effects across the entire population. (Sometimes for emphasis the ATE is called the *population average treatment effect*.) We can write the ATE parameter as

$$\tau_{ate} = \mathrm{E}[te_i] = \mathrm{E}[y_i(1) - y_i(0)] = \mathrm{E}[y_i(1)] - \mathrm{E}[y_i(0)], \tag{2.76}$$

where the final expression uses linearity of the expected value. Sometimes, to emphasize the population nature of τ_{ate} we write $\tau_{ate} = \mathrm{E}[y(1) - y(0)]$, where $[y(0), y(1)]$ are the two random variables representing the counterfactual outcomes in the population.

For each unit i let x_i be the program participation status—a binary variable. Then the observed outcome, y_i, can be written as

$$y_i = (1 - x_i)y_i(0) + x_i y_i(1), \tag{2.77}$$

which is just shorthand for $y_i = y_i(0)$ if $x_i = 0$ and $y_i = y_i(1)$ if $x_i = 1$. This equation precisely describes why, given a random sample from the population, we observe only one of $y_i(0)$ and $y_i(1)$.

To see how to estimate the average treatment effect, it is useful to rearrange (2.77):

$$y_i = y_i(0) + [y_i(1) - y_i(0)]x_i \qquad [2.78]$$

Now impose a simple (and, usually, unrealistic) constant treatment effect. Namely, for all *i,*

$$y_i(1) = \tau + y_i(0), \qquad [2.79]$$

or $\tau = y_i(1) - y_i(0)$. Plugging this into (2.78) gives

$$y_i = y_i(0) + \tau x_i.$$

Now write $y_i(0) = \alpha_0 + u_i(0)$ where, by definition, $\alpha_0 = \mathrm{E}[y_i(0)]$ and $\mathrm{E}[u_i(0)] = 0$. Plugging this in gives

$$y_i = \alpha_0 + \tau x_i + u_i(0). \qquad [2.80]$$

If we define $\beta_0 = \alpha_0$, $\beta_1 = \tau$, and $u_i = u_i(0)$ then the equation becomes exactly as in equation (2.48):

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where $\beta_1 = \tau$ is the treatment (or causal) effect.

We can easily determine that the simple regression estimator, which we now know is the difference in means estimator, is unbiased for the treatment effect, τ. If x_i is independent of $u_i(0)$ then

$$\mathrm{E}[u_i(0)|x_i] = 0,$$

so that SLR.4 holds. We have already shown that SLR.1 holds in our derivation of (2.80). As usual, we assume random sampling (SLR.2), and SLR.3 holds provided we have some treated units and some control units, a basic requirement. It is pretty clear we cannot learn anything about the effect of the intervention if all sampled units are in the control group or all are in the treatment group.

The assumption that x_i is independent of $u_i(0)$ is the same as x_i is independent of $y_i(0)$. This assumption can be guaranteed only under **random assignment**, whereby units are assigned to the treatment and control groups using a randomization mechanism that ignores any features of the individual units. For example, in evaluating a job training program, random assignment occurs if a coin is flipped to determine whether a worker is in the control group or treatment group. (The coin can be biased in the sense that the probability of a head need not be 0.5.) Random assignment can be compromised if units do not comply with their assignment.

Random assignment is the hallmark of a **randomized controlled trial (RCT)**, which has long been considered the gold standard for determing whether medical interventions have causal effects. RCTs generate the kind of experimental data of the type discussed in Chapter 1. In recent years, RCTs have become more popular in certain fields in economics, such as development economics and behavioral economics. Unfortunately, RCTs can be very expensive to implement, and in many cases randomizing subjects into control and treatment groups raises ethical issues. (For example, if giving low-income families access to free health care improves child health outcomes then randomizing some families into the control group means those children will have, on average, worse health outcomes than they could have otherwise.)

Even though RCTs are not always feasible for answering particular questions in economics and other fields, it is a good idea to think about the experiment one *would* run if random assignment were possible. Working through the simple thought experiment typically ensures that one is asking a sensible question before gathering nonexperimental data. For example, if we want to study the effects of Internet access in rural areas on student performance, we might not have the resources (or ethical clearance) to randomly assign Internet access to some students and not others. Nevertheless, thinking about how such an experiment would be implemented sharpens our thinking about the potential outcomes framework and what we mean by the treatment effect.

Our discussion of random assignment so far shows that, in the context of a constant treatment effect, the simple difference-in-means estimator, $\bar{y}_1 - \bar{y}_0$, is unbiased for τ. We can easily relax the constant treatment effect assumption. In general, the individual treatment effect can be written as

$$te_i = y_i(1) - y_i(0) = \tau_{ate} + [u_i(1) - u_i(0)], \qquad [2.81]$$

where $y_i(1) = \alpha_1 + u_i(1)$ and $\tau_{ate} = \alpha_1 - \alpha_0$. It is helpful to think of τ_{ate} as the average across the entire population and $u_i(1) - u_i(0)$ as the deviation from the population average for unit *i*. Plugging (2.81) into (2.78) gives

$$y_i = \alpha_0 + \tau_{ate}x_i + u_i(0) + [u_i(1) - u_i(0)]x_i \equiv \alpha_0 + \tau_{ate}x_i + u_i, \qquad [2.82]$$

where the error term is now

$$u_i = u_i(0) + [u_i(1) - u_i(0)]x_i.$$

The random assignment assumption is now that x_i is independent of $[u_i(0), u_i(1)]$. Even though u_i depends on x_i, the zero conditional mean assumption holds:

$$\begin{aligned} \mathrm{E}(u_i|x_i) &= \mathrm{E}[u_i(0)|x_i] + \mathrm{E}[u_i(1) - u_i(0)|x_i]x_i \\ &= 0 + 0 \cdot x_i = 0. \end{aligned}$$

We have again verified SLR.4, and so we conclude that the simple OLS estimator is unbiased for α_0 and τ_{ate},where $\hat{\tau}_{ate}$ is the difference-in-means estimator. [The error u_i is not independent of x_i. In particular, as shown in Problem 2.17, $\mathrm{Var}(u_i|x_i)$ differs across $x_i = 1$ and $x_i = 0$ if the variances of the potential outcomes differ. But remember, Assumption SLR.5 is not used to show the OLS estimators are unbiased.]

The fact that the simple regression estimator produces an unbiased estimator τ_{ate} when the treatment effects can vary arbitrarily across individual units is a very powerful result. However, it relies heavily on random assignment. Starting in Chapter 3, we will see how multiple regression analysis can be used when pure random assignment does not hold. Chapter 20, available as an online supplement, contains an accessible survey of advanced methods for estimating treatment effects.

EXAMPLE 2.14 Evaluating a Job Training Program

The data in JTRAIN2 are from an old, experimental job training program, where men with poor labor market histories were assigned to control and treatment groups. This data set has been used widely in the program evaluation literature to compare estimates from nonexperimental programs. The training assignment indicator is *train* and here we are interested in the outcome *re*78, which is (real) earnings in 1978 measured in thousands of dollars. Of the 445 men in the sample, 185 participated in the program in a period prior to 1978; the other 260 men comprise the control group.

The simple regression gives

$$\begin{aligned} \widehat{re78} &= 4.55 + 1.79\ train \\ n &= 445, R^2 = 0.018. \end{aligned}$$

From the earlier discussion, we know that 1.79 is the difference in average *re*78 between the treated and control groups, so men who participated in the program earned an average of $1,790 more than the men who did not. This is an economically large effect, as the dollars are 1978 dollars. Plus, the average earnings for men who did not participate is $4,550; in percentage terms, the gain in average earnings is about 39.3%, which is large. (We would need to know the costs of the program to do a benefit-cost analysis, but the benefits are nontrivial.)

Remember that the fundamental issue in program evaluation is that we do not observe any of the units in both states of the world. In this example, we only observe one of the two earnings outcomes for each men. Neverthless, random assignment into the treatment and control groups allows us to get an unbiased estimator of the average treatment effect.

Two final comments on this example. First, notice the very small *R*-squared: the training participation indicator explains less than two percent of the variation in *re*78 in the sample. We should not be surprised: many other factors, including education, experience, intelligence, age, motivation, and so on help determine labor market earnings. This is a good example to show how focusing on *R*-squared is not only unproductive, but it can be harmful. Beginning students sometimes think a small *R*-squared indicates "bias" in the OLS estimators. It does not. It simply means that the variance in the unobservables, $\mathrm{Var}(u)$, is large relative to $\mathrm{Var}(y)$. In this example, we know that Assumptions

SLR.1 to SLR.4 hold because of random assignment. Rightfully, none of these assumptions mentions how large *R*-squared must be; it is immaterial for the notion of unbiasedness.

A second comment is that, while the estimated economic effect of \$1,790 is large, we do not know whether this estimate is statistically significant. We will come to this topic in Chapter 4.

Before ending this chapter, it is important to head off possible confusion about two different ways the word "random" has been used in this subsection. First, the notion of random sampling is the one introduced in Assumption SLR.2 (and also discussed in Math Refresher C). Random sampling means that the data we obtain are independent, identically distributed draws from the population distribution represented by the random variables (x, y). It is important to understand that random sampling is a separate concept from random assignment, which means that x_i is determined independently of the counterfactuals $[y_i(0), y_i(1)]$. In Example 2.14, we obtained a random sample from the relevant population, and the assignment to treatment and control is randomized. But in other cases, random assignment will not hold even though we have random sampling. For example, it is relatively easy to draw a random sample from a large population of college-bound students and obtain outcomes on their SAT scores and whether they participated in an SAT preparation course. That does not mean that participation in a course is independent of the counterfactual outcomes. If we wanted to ensure independence between participation and the potential outcomes, we would randomly assign the students to take a course or not (and insist that students adhere to their assignments). If instead we obtain retrospective data—that is, we simply record whether a student has taken a preparation course—then the independence assumption underlying RA is unlikely to hold. But this has nothing to do with whether we obtained a random sample of students from the population. The general point is that Assumptions SLR.2 and SLR.4 are very different.

Summary

We have introduced the simple linear regression model in this chapter, and we have covered its basic properties. Given a random sample, the method of ordinary least squares is used to estimate the slope and intercept parameters in the population model. We have demonstrated the algebra of the OLS regression line, including computation of fitted values and residuals, and the obtaining of predicted changes in the dependent variable for a given change in the independent variable. In Section 2-4, we discussed two issues of practical importance: (1) the behavior of the OLS estimates when we change the units of measurement of the dependent variable or the independent variable and (2) the use of the natural log to allow for constant elasticity and constant semi-elasticity models.

In Section 2-5, we showed that, under the four Assumptions SLR.1 through SLR.4, the OLS estimators are unbiased. The key assumption is that the error term u has zero mean, or average, given any value of the independent variable x. Unfortunately, there are reasons to think this is false in many social science applications of simple regression, where the omitted factors in u are often correlated with x. When we add the assumption that the variance of the error given x is constant, we get simple formulas for the sampling variances of the OLS estimators. As we saw, the variance of the slope estimator $\hat{\beta}_1$ increases as the error variance increases, and it decreases when there is more sample variation in the independent variable. We also derived an unbiased estimator for $\sigma^2 = \text{Var}(u)$.

In Section 2-6, we briefly discussed regression through the origin, where the slope estimator is obtained under the assumption that the intercept is zero. Sometimes, this is useful, but it appears infrequently in applied work.

In Section 2-7 we covered the important case where x is a binary variable, and showed that the OLS "slope" estimate is simply $\hat{\beta}_0 = \bar{y}_1 - \bar{y}_0$, the difference in the averages of y_i between the $x_i = 1$ and $x_i = 0$ subsamples. We also discussed how, in the context of causal inference, $\hat{\beta}_1$ is an unbiased estimator of the average treatment effect under random assignment into the control and treatment groups. In Chapter 3 and beyond, we will study the case where the intervention or treatment is not randomized, but depends on observed and even unobserved factors.

Much work is left to be done. For example, we still do not know how to test hypotheses about the population parameters, β_0 and β_1. Thus, although we know that OLS is unbiased for the population parameters

under Assumptions SLR.1 through SLR.4, we have no way of drawing inferences about the population. Other topics, such as the efficiency of OLS relative to other possible procedures, have also been omitted.

The issues of confidence intervals, hypothesis testing, and efficiency are central to multiple regression analysis as well. Because the way we construct confidence intervals and test statistics is very similar for multiple regression—and because simple regression is a special case of multiple regression—our time is better spent moving on to multiple regression, which is much more widely applicable than simple regression. Our purpose in Chapter 2 was to get you thinking about the issues that arise in econometric analysis in a fairly simple setting.

THE GAUSS-MARKOV ASSUMPTIONS FOR SIMPLE REGRESSION

For convenience, we summarize the **Gauss-Markov assumptions** that we used in this chapter. It is important to remember that only SLR.1 through SLR.4 are needed to show $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. We added the homoskedasticity assumption, SLR.5, to obtain the usual OLS variance formulas (2.57) and (2.58).

Assumption SLR.1 (Linear in Parameters)

In the population model, the dependent variable, y, is related to the independent variable, x, and the error (or disturbance), u, as

$$y = \beta_0 + \beta_1 x + u,$$

where β_0 and β_1 are the population intercept and slope parameters, respectively.

Assumption SLR.2 (Random Sampling)

We have a random sample of size n, $\{(x_i,y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption SLR.1.

Assumption SLR.3 (Sample Variation in the Explanatory Variable)

The sample outcomes on x, namely, $\{x_i, i = 1, \ldots, n\}$, are not all the same value.

Assumption SLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any value of the explanatory variable. In other words,

$$\mathrm{E}(u|x) = 0.$$

Assumption SLR.5 (Homoskedasticity)

The error u has the same variance given any value of the explanatory variable. In other words,

$$\mathrm{Var}(u|x) = \sigma^2.$$

Key Terms

Average Treatment Effect (ATE)
Average Causal Effect (ACE)
Binary (Dummy) Variable
Causal (Treatment) Effect
Coefficient of Determination
Constant Elasticity Model
Control Group
Control Variable
Covariate
Degrees of Freedom
Dependent Variable
Elasticity
Error Term (Disturbance)
Error Variance
Explained Sum of Squares (SSE)
Explained Variable
Explanatory Variable
First Order Conditions
Fitted Value
Gauss-Markov Assumptions
Heteroskedasticity
Homoskedasticity
Independent Variable
Intercept Parameter
Mean Independent
OLS Regression Line
Ordinary Least Squares (OLS)
Population Regression Function (PRF)
Predicted Variable
Predictor Variable
Random Assignment
Randomized Controlled Trial (RCT)
Regressand
Regression through the Origin
Regressor
Residual
Residual Sum of Squares (SSR)
Response Variable
R-squared
Sample Regression Function (SRF)
Semi-elasticity
Simple Linear Regression Model
Slope Parameter
Standard Error of $\hat{\beta}_1$
Standard Error of the Regression (SER)
Sum of Squared Residuals
Total Sum of Squares (SST)
Treatment Group
Zero Conditional Mean Assumption

Problems

1 Let *kids* denote the number of children ever born to a woman, and let *educ* denote years of education for the woman. A simple model relating fertility to years of education is

$$kids = \beta_0 + \beta_1 educ + u,$$

where u is the unobserved error.

(i) What kinds of factors are contained in u? Are these likely to be correlated with level of education?

(ii) Will a simple regression analysis uncover the ceteris paribus effect of education on fertility? Explain.

2 In the simple linear regression model $y = \beta_0 + \beta_1 x + u$, suppose that $\mathrm{E}(u) \neq 0$. Letting $\alpha_0 = \mathrm{E}(u)$, show that the model can always be rewritten with the same slope, but a new intercept and error, where the new error has a zero expected value.

3 The following table contains the *ACT* scores and the *GPA* (grade point average) for eight college students. Grade point average is based on a four-point scale and has been rounded to one digit after the decimal.

Student	*GPA*	*ACT*
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

(i) Estimate the relationship between *GPA* and *ACT* using OLS; that is, obtain the intercept and slope estimates in the equation

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT.$$

Comment on the direction of the relationship. Does the intercept have a useful interpretation here? Explain. How much higher is the *GPA* predicted to be if the *ACT* score is increased by five points?

(ii) Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum to zero.

(iii) What is the predicted value of *GPA* when $ACT = 20$?

(iv) How much of the variation in *GPA* for these eight students is explained by *ACT*? Explain.

4 The data set BWGHT contains data on births to women in the United States. Two variables of interest are the dependent variable, infant birth weight in ounces (*bwght*), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy (*cigs*). The following simple regression was estimated using data on $n = 1{,}388$ *births*:

$$\widehat{bwght} = 119.77 - 0.514\, cigs$$

(i) What is the predicted birth weight when $cigs = 0$? What about when $cigs = 20$ (one pack per day)? Comment on the difference.

(ii) Does this simple regression necessarily capture a causal relationship between the child's birth weight and the mother's smoking habits? Explain.

(iii) To predict a birth weight of 125 ounces, what would *cigs* have to be? Comment.

(iv) The proportion of women in the sample who do not smoke while pregnant is about .85. Does this help reconcile your finding from part (iii)?

5 In the linear consumption function

$$\widehat{cons} = \hat{\beta}_0 + \hat{\beta}_1 inc,$$

the (estimated) *marginal propensity to consume* (MPC) out of income is simply the slope, $\hat{\beta}_1$, while the *average propensity to consume* (APC) is $\widehat{cons}/inc = \hat{\beta}_0/inc + \hat{\beta}_1$. Using observations for 100 families on annual income and consumption (both measured in dollars), the following equation is obtained:

$$\widehat{cons} = -124.84 + 0.853\, inc$$

$$n = 100, R^2 = 0.692.$$

(i) Interpret the intercept in this equation, and comment on its sign and magnitude.

(ii) What is the predicted consumption when family income is $30,000?

(iii) With *inc* on the *x*-axis, draw a graph of the estimated MPC and APC.

6 Using data from 1988 for houses sold in Andover, Massachusetts, from Kiel and McClain (1995), the following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*dist*):

$$\widehat{\log(price)} = 9.40 + 0.312 \log(dist)$$

$$n = 135, R^2 = 0.162.$$

(i) Interpret the coefficient on log(*dist*). Is the sign of this estimate what you expect it to be?

(ii) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of *price* with respect to *dist*? (Think about the city's decision on where to put the incinerator.)

(iii) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

7 Consider the savings function

$$sav = \beta_0 + \beta_1 inc + u, u = \sqrt{inc} \cdot e,$$

where e is a random variable with $\mathrm{E}(e) = 0$ and $\mathrm{Var}(e) = \sigma_e^2$. Assume that e is independent of *inc*.

(i) Show that $\mathrm{E}(u|inc) = 0$, so that the key zero conditional mean assumption (Assumption SLR.4) is satisfied. [*Hint:* If e is independent of *inc*, then $\mathrm{E}(e|inc) = \mathrm{E}(e)$.]

(ii) Show that $\mathrm{Var}(u|inc) = \sigma_e^2 inc$, so that the homoskedasticity Assumption SLR.5 is violated. In particular, the variance of *sav* increases with *inc*. [*Hint:* $\mathrm{Var}(e|inc) = \mathrm{Var}(e)$ if e and *inc* are independent.]

(iii) Provide a discussion that supports the assumption that the variance of savings increases with family income.

8 Consider the standard simple regression model $y = \beta_0 + \beta_1 x + u$ under the Gauss-Markov Assumptions SLR.1 through SLR.5. The usual OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for their respective population parameters. Let $\tilde{\beta}_1$ be the estimator of β_1 obtained by assuming the intercept is zero (see Section 2-6).

(i) Find $\mathrm{E}(\tilde{\beta}_1)$ in terms of the x_i, β_0, and β_1. Verify that $\tilde{\beta}_1$ is unbiased for β_1 when the population intercept (β_0) is zero. Are there other cases where $\tilde{\beta}_1$ is unbiased?

(ii) Find the variance of $\tilde{\beta}_1$. (*Hint:* The variance does not depend on β_0.)

(iii) Show that $\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1)$. [*Hint:* For any sample of data, $\sum_{i-1}^{n} x_i^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2$, with strict inequality unless $\bar{x} = 0$.]

(iv) Comment on the tradeoff between bias and variance when choosing between $\hat{\beta}_1$ and $\tilde{\beta}_1$.

9 (i) Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the intercept and slope from the regression of y_i on x_i, using n observations. Let c_1 and c_2, with $c_2 \neq 0$, be constants. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of $c_1 y_i$ on $c_2 x_i$. Show that $\tilde{\beta}_1 = (c_1/c_2)\hat{\beta}_0$ and $\tilde{\beta}_0 = c_1\hat{\beta}_0$, thereby verifying the claims on units of measurement in Section 2-4. [*Hint:* To obtain $\tilde{\beta}_1$, plug the scaled versions of x and y into (2.19). Then, use (2.17) for $\tilde{\beta}_0$, being sure to plug in the scaled x and y *and* the correct slope.]

(ii) Now, let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be from the regression of $(c_1 + y_i)$ on $(c_2 + x_i)$ (with no restriction on c_1 or c_2). Show that $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_0 = \hat{\beta}_0 + c_1 - c_2\hat{\beta}_1$.

(iii) Now, let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS estimates from the regression $\log(y_i)$ on x_i, where we must assume $y_i > 0$ for all i. For $c_1 > 0$, let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of $\log(c_1 y_i)$ on x_i. Show that $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_0 = \log(c_1) + \hat{\beta}_0$.

(iv) Now, assuming that $x_i > 0$ for all i, let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of y_i on $\log(c_2 x_i)$. How do $\tilde{\beta}_0$ and $\tilde{\beta}_1$ compare with the intercept and slope from the regression of y_i on $\log(x_i)$?

10 Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS intercept and slope estimators, respectively, and let $\bar{u}$ be the sample average of the errors (not the residuals!).

(i) Show that $\hat{\beta}_1$ can be written as $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$, where $w_i = d_i/\text{SST}_x$ and $d_i = x_i - \bar{x}$.

(ii) Use part (i), along with $\sum_{i=1}^{n} w_i = 0$, to show that $\hat{\beta}_1$ and $\bar{u}$ are uncorrelated. [*Hint*: You are being asked to show that $\text{E}[(\hat{\beta}_1 - \beta_1) \cdot \bar{u}] = 0$.]

(iii) Show that $\hat{\beta}_0$ can be written as $\hat{\beta}_0 = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}$.

(iv) Use parts (ii) and (iii) to show that $\text{Var}(\hat{\beta}_0) = \sigma^2/n + \sigma^2(\bar{x})^2/\text{SST}_x$.

(v) Do the algebra to simplify the expression in part (iv) to equation (2.58). [*Hint*: $\text{SST}_x/n = n^{-1}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$.]

11 Suppose you are interested in estimating the effect of hours spent in an SAT preparation course (*hours*) on total SAT score (*sat*). The population is all college-bound high school seniors for a particular year.

(i) Suppose you are given a grant to run a controlled experiment. Explain how you would structure the experiment in order to estimate the causal effect of *hours* on *sat*.

(ii) Consider the more realistic case where students choose how much time to spend in a preparation course, and you can only randomly sample *sat* and *hours* from the population. Write the population model as

$$sat = \beta_0 + \beta_1 hours + u$$

where, as usual in a model with an intercept, we can assume $\text{E}(u) = 0$. List at least two factors contained in u. Are these likely to have positive or negative correlation with *hours*?

(iii) In the equation from part (ii), what should be the sign of β_1 if the preparation course is effective?

(iv) In the equation from part (ii), what is the interpretation of β_0?

12 Consider the problem described at the end of Section 2-6, running a regression and only estimating an intercept.

(i) Given a sample $\{y_i: i = 1, 2, \ldots, n\}$, let $\tilde{\beta}_0$ be the solution to

$$\min_{b_0} \sum_{i=1}^{n} (y_i - b_0)^2.$$

Show that $\tilde{\beta}_0 = \bar{y}$, that is, the sample average minimizes the sum of squared residuals. (*Hint*: You may use one-variable calculus or you can show the result directly by adding and subtracting $\bar{y}$ inside the squared residual and then doing a little algebra.)

(ii) Define residuals $\tilde{u}_i = y_i - \bar{y}$. Argue that these residuals always sum to zero.

13 Let y be any response variable and x a binary explanatory variable. Let $\{(x_i, y_i): i = 1, \ldots, n\}$ be a sample of size n. Let n_0 be the number of observations with $x_i = 0$ and n_1 the number of observations with $x_i = 1$. Let $\bar{y}_0$ be the average of the y_i with $x_i = 0$ and $\bar{y}_1$ the average of the y_i with $x_i = 1$.

(i) Explain why we can write

$$n_0 = \sum_{i=1}^{n}(1 - x_i), n_1 = \sum_{i=1}^{n}x_i.$$

Show that $\bar{x} = n_1/n$ and $(1 - x) = n_0/n$. How do you interpret $\bar{x}$?

(ii) Argue that

$$\bar{y}_0 = n_0^{-1}\sum_{i=1}^{n}(1 - x_i)y_i, \bar{y}_1 = n_1^{-1}\sum_{i=1}^{n}x_i y_i.$$

(iii) Show that the average of y_i in the entire sample, $\bar{y}$, can be written as a weighted average:

$$\bar{y} = (1 - \bar{x})\bar{y}_0 + \bar{x}\bar{y}_1.$$

[*Hint*: Write $y_i = (1 - x_i)y_i + x_i y_i$.]

(iv) Show that when x_i is binary,

$$n^{-1}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2 = \bar{x}(1 - \bar{x}).$$

[*Hint*: When x_i is binary, $x_i^2 = x_i$.]

(v) Show that

$$n^{-1}\sum_{i=1}^{n}x_i y_i - \bar{x}\bar{y} = \bar{x}(1 - \bar{x})(\bar{y}_1 - \bar{y}_0).$$

(vi) Use parts (iv) and (v) to obtain (2.74).

(vii) Derive equation (2.73).

14 In the context of Problem 2.13, suppose y_i is also binary. For concreteness, y_i indicates whether worker i is employed after a job training program, where $y_i = 1$ means has a job, $y_i = 0$ means does not have a job. Here, x_i indicates participation in the job training program. Argue that $\hat{\beta}_1$ is the difference in employment rates between those who participated in the program and those who did not.

15 Consider the potential outcomes framework from Section 2.7a, where $y_i(0)$ and $y_i(1)$ are the potential outcomes in each treatment state.

(i) Show that if we could observe $y_i(0)$ and $y_i(1)$ for all i then an unbiased estimator of τ_{ate} would be

$$n^{-1}\sum_{i=1}^{n}[y_i(1) - y_i(0)] = \bar{y}(1) - \bar{y}(0).$$

This is sometimes called the *sample average treatment effect.*

(ii) Explain why the observed sample averages, $\bar{y}_0$ and $\bar{y}_1$, are not the same as $\bar{y}(0)$ and $\bar{y}(1)$, respectively, by writing $\bar{y}_0$ and $\bar{y}_1$ in terms of $y_i(0)$ and $y_i(1)$, respectively.

16 In the potential outcomes framework, suppose that program *eligibility* is randomly assigned but participation cannot be enforced. To formally describe this situation, for each person i, z_i is the eligibility indicator and x_i is the participation indicator. Randomized eligibility means z_i is independent of $[y_i(0), y_i(1)]$ but x_i might not satisfy the independence assumption.

(i) Explain why the difference in means estimator is generally no longer unbiased.

(ii) In the context of a job training program, what kind of individual behavior would cause bias?

17 In the potential outcomes framework with heterogeneous (nonconstant) treatment effect, write the error as

$$u_i = (1 - x_i)u_i(0) + x_i u_i(1).$$

Let $\sigma_0^2 = \text{Var}[u_i(0)]$ and $\sigma_1^2 = \text{Var}[u_i(1)]$. Assume random assignment.

(i) Find $\text{Var}(u_i|x_i)$.

(ii) When is $\text{Var}(u_i|x_i)$ constant?

18 Let x be a binary explanatory variable and suppose $P(x = 1) = \rho$ for $0 < \rho < 1$.

(i) If you draw a random sample of size n, find the probability—call it γ_n—that Assumption SLR.3 fails. [Hint: Find the probability of observing all zeros or all ones for the x_i.] Argue that $\gamma_n \to 0$ as $n \to \infty$.

(ii) If $\rho = 0.5$, compute the probablity in part (i) for $n = 10$ and $n = 100$. Discuss.

(iii) Do the calculations from part (ii) with $\rho = 0.9$. How do your answers compare with part (ii)?

Computer Exercises

C1 The data in 401K are a subset of data analyzed by Papke (1995) to study the relationship between participation in a 401(k) pension plan and the generosity of the plan. The variable *prate* is the percentage of eligible workers with an active account; this is the variable we would like to explain. The measure of generosity is the plan match rate, *mrate*. This variable gives the average amount the firm contributes to each worker's plan for each \$1 contribution by the worker. For example, if $mrate = 0.50$, then a \$1 contribution by the worker is matched by a 50¢ contribution by the firm.

(i) Find the average participation rate and the average match rate in the sample of plans.

(ii) Now, estimate the simple regression equation

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate,$$

and report the results along with the sample size and R-squared.

(iii) Interpret the intercept in your equation. Interpret the coefficient on *mrate*.

(iv) Find the predicted *prate* when $mrate = 3.5$. Is this a reasonable prediction? Explain what is happening here.

(v) How much of the variation in *prate* is explained by *mrate*? Is this a lot in your opinion?

C2 The data set in CEOSAL2 contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollars, and *ceoten* is prior number of years as company CEO.

(i) Find the average salary and the average tenure in the sample.

(ii) How many CEOs are in their first year as CEO (that is, $ceoten = 0$)? What is the longest tenure as a CEO?

(iii) Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten + u,$$

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

C3 Use the data in SLEEP75 from Biddle and Hamermesh (1990) to study whether there is a tradeoff between the time spent sleeping per week and the time spent in paid work. We could use either variable as the dependent variable. For concreteness, estimate the model

$$sleep = \beta_0 + \beta_1 totwrk + u,$$

where *sleep* is minutes spent sleeping at night per week and *totwrk* is total minutes worked during the week.

(i) Report your results in equation form along with the number of observations and R^2. What does the intercept in this equation mean?

(ii) If *totwrk* increases by 2 hours, by how much is *sleep* estimated to fall? Do you find this to be a large effect?

C4 Use the data in WAGE2 to estimate a simple regression explaining monthly salary (*wage*) in terms of IQ score (*IQ*).

(i) Find the average salary and average IQ in the sample. What is the sample standard deviation of IQ? (IQ scores are standardized so that the average in the population is 100 with a standard deviation equal to 15.)

(ii) Estimate a simple regression model where a one-point increase in *IQ* changes *wage* by a constant dollar amount. Use this model to find the predicted increase in wage for an increase in *IQ* of 15 points. Does *IQ* explain most of the variation in *wage*?

(iii) Now, estimate a model where each one-point increase in *IQ* has the same percentage effect on *wage*. If *IQ* increases by 15 points, what is the approximate percentage increase in predicted *wage*?

C5 For the population of firms in the chemical industry, let *rd* denote annual expenditures on research and development, and let *sales* denote annual sales (both are in millions of dollars).

(i) Write down a model (not an estimated equation) that implies a constant elasticity between *rd* and *sales*. Which parameter is the elasticity?

(ii) Now, estimate the model using the data in RDCHEM. Write out the estimated equation in the usual form. What is the estimated elasticity of *rd* with respect to *sales*? Explain in words what this elasticity means.

C6 We used the data in MEAP93 for Example 2.12. Now we want to explore the relationship between the math pass rate (*math10*) and spending per student (*expend*).

(i) Do you think each additional dollar spent has the same effect on the pass rate, or does a diminishing effect seem more appropriate? Explain.

(ii) In the population model

$$math10 = \beta_0 + \beta_1 \log(expend) + u,$$

argue that $\beta_1/10$ is the percentage point change in *math10* given a 10% increase in *expend*.

(iii) Use the data in MEAP93 to estimate the model from part (ii). Report the estimated equation in the usual way, including the sample size and *R*-squared.

(iv) How big is the estimated spending effect? Namely, if spending increases by 10%, what is the estimated percentage point increase in *math10*?

(v) One might worry that regression analysis can produce fitted values for *math10* that are greater than 100. Why is this not much of a worry in this data set?

C7 Use the data in CHARITY [obtained from Franses and Paap (2001)] to answer the following questions:

(i) What is the average gift in the sample of 4,268 people (in Dutch guilders)? What percentage of people gave no gift?

(ii) What is the average mailings per year? What are the minimum and maximum values?

(iii) Estimate the model

$$gift = \beta_0 + \beta_1 mailsyear + u$$

by OLS and report the results in the usual way, including the sample size and *R*-squared.

(iv) Interpret the slope coefficient. If each mailing costs one guilder, is the charity expected to make a net gain on each mailing? Does this mean the charity makes a net gain on every mailing? Explain.

(v) What is the smallest predicted charitable contribution in the sample? Using this simple regression analysis, can you ever predict zero for *gift*?

C8 To complete this exercise you need a software package that allows you to generate data from the uniform and normal distributions.

(i) Start by generating 500 observations on x_i—the explanatory variable—from the uniform distribution with range [0,10]. (Most statistical packages have a command for the Uniform(0,1) distribution; just multiply those observations by 10.) What are the sample mean and sample standard deviation of the x_i?

(ii) Randomly generate 500 errors, u_i, from the Normal(0,36) distribution. (If you generate a Normal(0,1), as is commonly available, simply multiply the outcomes by six.) Is the sample average of the u_i exactly zero? Why or why not? What is the sample standard deviation of the u_i?

(iii) Now generate the y_i as

$$y_i = 1 + 2x_i + u_i \equiv \beta_0 + \beta_1 x_i + u_i;$$

that is, the population intercept is one and the population slope is two. Use the data to run the regression of y_i on x_i. What are your estimates of the intercept and slope? Are they equal to the population values in the above equation? Explain.

(iv) Obtain the OLS residuals, $\hat{u}_i$, and verify that equation (2.60) holds (subject to rounding error).

(v) Compute the same quantities in equation (2.60) but use the errors u_i in place of the residuals. Now what do you conclude?

(vi) Repeat parts (i), (ii), and (iii) with a new sample of data, starting with generating the x_i. Now what do you obtain for $\hat{\beta}_0$ and $\hat{\beta}_1$? Why are these different from what you obtained in part (iii)?

C9 Use the data in COUNTYMURDERS to answer these questions. Use only the data for 1996.

(i) How many counties had zero murders in 1996? How many counties had at least one execution? What is the largest number of executions?

(ii) Estimate the equation

$$murders = \beta_0 + \beta_1 execs + u$$

by OLS and report the results in the usual way, including sample size and R-squared.

(iii) Interpret the slope coefficient reported in part (ii). Does the estimated equation suggest a deterrent effect of capital punishment?

(iv) What is the smallest number of murders that can be predicted by the equation? What is the residual for a county with zero executions and zero murders?

(v) Explain why a simple regression analysis is not well suited for determining whether capital punishment has a deterrent effect on murders.

C10 The data set in CATHOLIC includes test score information on over 7,000 students in the United States who were in eighth grade in 1988. The variables *math*12 and *read*12 are scores on twelfth grade standardized math and reading tests, respectively.

(i) How many students are in the sample? Find the means and standard deviations of *math*12 and *read*12.

(ii) Run the simple regression of *math*12 on *read*12 to obtain the OLS intercept and slope estimates. Report the results in the form

$$\widehat{math12} = \hat{\beta}_0 + \hat{\beta}_1 read12$$

$$n = ?, R^2 = ?$$

where you fill in the values for $\hat{\beta}_0$ and $\hat{\beta}_1$ and also replace the question marks.

(iii) Does the intercept reported in part (ii) have a meaningful interpretation? Explain.

(iv) Are you surprised by the $\hat{\beta}_1$ that you found? What about R^2?

(v) Suppose that you present your findings to a superintendent of a school district, and the superintendent says, "Your findings show that to improve math scores we just need to improve reading scores, so we should hire more reading tutors." How would you respond to this comment? (Hint: If you instead run the regression of *read*12 on *math*12, what would you expect to find?)

C11 Use the data in GPA1 to answer these questions. It is a sample of Michigan State University undergraduates from the mid-1990s, and includes current college GPA, *colGPA*, and a binary variable indicating whether the student owned a personal computer (*PC*).

(i) How many students are in the sample? Find the average and highest college GPAs.

(ii) How many students owned their own PC?

(iii) Estimate the simple regression equation

$$colGPA = \beta_0 + \beta_1 PC + u$$

and report your estimates for β_0 and β_1. Interpret these estimates, including a discussion ofthe magnitudes.

(iv) What is the *R*-squared from the regression? What do you make of its magnitude?

(v) Does your finding in part (iii) imply that owning a PC has a causal effect on *colGPA*? Explain.

APPENDIX 2A

Minimizing the Sum of Squared Residuals

We show that the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ do minimize the sum of squared residuals, as asserted in Section 2-2. Formally, the problem is to characterize the solutions $\hat{\beta}_0$ and $\hat{\beta}_1$ to the minimization problem

$$\min_{b_0,b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

where b_0 and b_1 are the dummy arguments for the optimization problem; for simplicity, call this function $Q(b_0, b_1)$. By a fundamental result from multivariable calculus (see Math Refresher A), a necessary condition for $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the minimization problem is that the partial derivatives of $Q(b_0, b_1)$ with respect to b_0 and b_1 must be zero when evaluated at $\hat{\beta}_0, \hat{\beta}_1$: $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_0 = 0$ and $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_1 = 0$. Using the chain rule from calculus, these two equations become

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$-2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

These two equations are just (2.14) and (2.15) multiplied by $-2n$ and, therefore, are solved by the same $\hat{\beta}_0$ and $\hat{\beta}_1$.

How do we know that we have actually minimized the sum of squared residuals? The first order conditions are necessary but not sufficient conditions. One way to verify that we have minimized the sum of squared residuals is to write, for any b_0 and b_1,

$$\begin{aligned} Q(b_0, b_1) &= \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2 \\ &= \sum_{i=1}^{n} [\hat{u}_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2 \\ &= \sum_{i=1}^{n} \hat{u}_i^2 + n(\hat{\beta}_0 - b_0)^2 + (\hat{\beta}_1 - b_1)^2 \sum_{i=1}^{n} x_i^2 + 2(\hat{\beta}_0 - b_0)(\hat{\beta}_1 - b_1)\sum_{i=1}^{n} x_i, \end{aligned}$$

where we have used equations (2.30) and (2.31). The first term does not depend on b_0 or b_1, while the sum of the last three terms can be written as

$$\sum_{i=1}^{n} [(\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1)x_i]^2,$$

as can be verified by straightforward algebra. Because this is a sum of squared terms, the smallest it can be is zero. Therefore, it is smallest when $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$.

CHAPTER 3

Multiple Regression Analysis: Estimation

In Chapter 2, we learned how to use simple regression analysis to explain a dependent variable, y, as a function of a single independent variable, x. The primary drawback in using simple regression analysis for empirical work is that it is very difficult to draw ceteris paribus conclusions about how x affects y: the key assumption, SLR.4—that all other factors affecting y are uncorrelated with x—is often unrealistic.

Multiple regression analysis is more amenable to ceteris paribus analysis because it allows us to *explicitly* control for many other factors that simultaneously affect the dependent variable. This is important both for testing economic theories and for evaluating policy effects when we must rely on nonexperimental data. Because multiple regression models can accommodate many explanatory variables that may be correlated, we can hope to infer causality in cases where simple regression analysis would be misleading.

Naturally, if we add more factors to our model that are useful for explaining y, then more of the variation in y can be explained. Thus, multiple regression analysis can be used to build better models for predicting the dependent variable.

An additional advantage of multiple regression analysis is that it can incorporate fairly general functional form relationships. In the simple regression model, only one function of a single explanatory variable can appear in the equation. As we will see, the multiple regression model allows for much more flexibility.

Section 3-1 formally introduces the multiple regression model and further discusses the advantages of multiple regression over simple regression. In Section 3-2, we demonstrate how to estimate the parameters in the multiple regression model using the method of ordinary least squares.

In Sections 3-3, 3-4, and 3-5, we describe various statistical properties of the OLS estimators, including unbiasedness and efficiency.

The multiple regression model is still the most widely used vehicle for empirical analysis in economics and other social sciences. Likewise, the method of ordinary least squares is popularly used for estimating the parameters of the multiple regression model.

3-1 Motivation for Multiple Regression

3-1a The Model with Two Independent Variables

We begin with some simple examples to show how multiple regression analysis can be used to solve problems that cannot be solved by simple regression.

The first example is a simple variation of the wage equation introduced in Chapter 2 for obtaining the effect of education on hourly wage:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u, \quad [3.1]$$

where *exper* is years of labor market experience. Thus, *wage* is determined by the two explanatory or independent variables, education and experience, and by other unobserved factors, which are contained in u. We are still primarily interested in the effect of *educ* on *wage*, holding fixed all other factors affecting *wage*; that is, we are interested in the parameter β_1.

Compared with a simple regression analysis relating *wage* to *educ*, equation (3.1) effectively takes *exper* out of the error term and puts it explicitly in the equation. Because *exper* appears in the equation, its coefficient, β_2, measures the ceteris paribus effect of *exper* on *wage*, which is also of some interest.

Not surprisingly, just as with simple regression, we will have to make assumptions about how u in (3.1) is related to the independent variables, *educ* and *exper*. However, as we will see in Section 3-2, there is one thing of which we can be confident: because (3.1) contains experience explicitly, we will be able to measure the effect of education on wage, holding experience fixed. In a simple regression analysis—which puts *exper* in the error term—we would have to assume that experience is uncorrelated with education, a tenuous assumption.

As a second example, consider the problem of explaining the effect of per-student spending (*expend*) on the average standardized test score (*avgscore*) at the high school level. Suppose that the average test score depends on funding, average family income (*avginc*), and other unobserved factors:

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u. \quad [3.2]$$

The coefficient of interest for policy purposes is β_1, the ceteris paribus effect of *expend* on *avgscore*. By including *avginc* explicitly in the model, we are able to control for its effect on *avgscore*. This is likely to be important because average family income tends to be correlated with per-student spending: spending levels are often determined by both property and local income taxes. In simple regression analysis, *avginc* would be included in the error term, which would likely be correlated with *expend*, causing the OLS estimator of β_1 in the two-variable model to be biased.

In the two previous similar examples, we have shown how observable factors other than the variable of primary interest [*educ* in equation (3.1) and *expend* in equation (3.2)] can be included in a regression model. Generally, we can write a model with two independent variables as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad [3.3]$$

where

β_0 is the intercept.
β_1 measures the change in y with respect to x_1, holding other factors fixed.
β_2 measures the change in y with respect to x_2, holding other factors fixed.

Multiple regression analysis is also useful for generalizing functional relationships between variables. As an example, suppose family consumption (*cons*) is a quadratic function of family income (*inc*):

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u, \qquad [3.4]$$

where u contains other factors affecting consumption. In this model, consumption depends on only one observed factor, income; so it might seem that it can be handled in a simple regression framework. But the model falls outside simple regression because it contains two functions of income, *inc* and inc^2 (and therefore three parameters, β_0, β_1, and β_2). Nevertheless, the consumption function is easily written as a regression model with two independent variables by letting $x_1 = inc$ and $x_2 = inc^2$.

Mechanically, there will be *no* difference in using the method of ordinary least squares (introduced in Section 3-2) to estimate equations as different as (3.1) and (3.4). Each equation can be written as (3.3), which is all that matters for computation. There is, however, an important difference in how one *interprets* the parameters. In equation (3.1), β_1 is the ceteris paribus effect of *educ* on *wage*. The parameter β_1 has no such interpretation in (3.4). In other words, it makes no sense to measure the effect of *inc* on *cons* while holding inc^2 fixed, because if *inc* changes, then so must inc^2! Instead, the change in consumption with respect to the change in income—the marginal propensity to consume—is approximated by

$$\frac{\Delta cons}{\Delta inc} \approx \beta_1 + 2\beta_2 inc.$$

See Math Refresher A for the calculus needed to derive this equation. In other words, the marginal effect of income on consumption depends on β_2 as well as on β_1 and the level of income. This example shows that, in any particular application, the definitions of the independent variables are crucial. But for the theoretical development of multiple regression, we can be vague about such details. We will study examples like this more completely in Chapter 6.

In the model with two independent variables, the key assumption about how u is related to x_1 and x_2 is

$$\mathrm{E}(u|x_1, x_2) = 0. \qquad [3.5]$$

The interpretation of condition (3.5) is similar to the interpretation of Assumption SLR.4 for simple regression analysis. It means that, for any values of x_1 and x_2 in the population, the average of the unobserved factors is equal to zero. As with simple regression, the important part of the assumption is that the expected value of u is the same for all combinations of x_1 and x_2; that this common value is zero is no assumption at all as long as the intercept β_0 is included in the model (see Section 2-1).

How can we interpret the zero conditional mean assumption in the previous examples? In equation (3.1), the assumption is $\mathrm{E}(u|educ,exper) = 0$. This implies that other factors affecting *wage* are not related on average to *educ* and *exper*. Therefore, if we think innate ability is part of u, then we will need average ability levels to be the same across all combinations of education and experience in the working population. This may or may not be true, but, as we will see in Section 3-3, this is the question we need to ask in order to determine whether the method of ordinary least squares produces unbiased estimators.

The example measuring student performance [equation (3.2)] is similar to the wage equation. The zero conditional mean assumption is $\mathrm{E}(u|expend,avginc) = 0$, which means that other

GOING FURTHER 3.1

A simple model to explain city murder rates (*murdrate*) in terms of the probability of conviction (*prbconv*) and average sentence length (*avgsen*) is

$$murdrate = \beta_0 + \beta_1 prbconv + \beta_2 avgsen + u.$$

What are some factors contained in u? Do you think the key assumption (3.5) is likely to hold?

factors affecting test scores—school or student characteristics—are, on average, unrelated to per-student funding and average family income.

When applied to the quadratic consumption function in (3.4), the zero conditional mean assumption has a slightly different interpretation. Written literally, equation (3.5) becomes $E(u|inc,inc^2) = 0$. Because inc^2 is known when inc is known, including inc^2 in the expectation is redundant: $E(u|inc,inc^2) = 0$ is the same as $E(u|inc) = 0$. Nothing is wrong with putting inc^2 along with inc in the expectation when stating the assumption, but $E(u|inc) = 0$ is more concise.

3-1b The Model with *k* Independent Variables

Once we are in the context of multiple regression, there is no need to stop with two independent variables. Multiple regression analysis allows many observed factors to affect y. In the wage example, we might also include amount of job training, years of tenure with the current employer, measures of ability, and even demographic variables like the number of siblings or mother's education. In the school funding example, additional variables might include measures of teacher quality and school size.

The general **multiple linear regression (MLR) model** (also called the *multiple regression model*) can be written in the population as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u, \qquad [3.6]$$

where

β_0 is the **intercept**.
β_1 is the parameter associated with x_1.
β_2 is the parameter associated with x_2, and so on.

Because there are k independent variables and an intercept, equation (3.6) contains $k + 1$ (unknown) population parameters. For shorthand purposes, we will sometimes refer to the parameters other than the intercept as **slope parameters**, even though this is not always literally what they are. [See equation (3.4), where neither β_1 nor β_2 is itself a slope, but together they determine the slope of the relationship between consumption and income.]

The terminology for multiple regression is similar to that for simple regression and is given in Table 3.1. Just as in simple regression, the variable u is the **error term** or **disturbance**. It contains factors other than $x_1, x_2, \ldots, x_k$ that affect y. No matter how many explanatory variables we include in our model, there will always be factors we cannot include, and these are collectively contained in u.

When applying the general multiple regression model, we must know how to interpret the parameters. We will get plenty of practice now and in subsequent chapters, but it is useful at this point to be reminded of some things we already know. Suppose that CEO salary (*salary*) is related to firm sales (*sales*) and CEO tenure (*ceoten*) with the firm by

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u. \qquad [3.7]$$

This fits into the multiple regression model (with $k = 3$) by defining $y = \log(salary)$, $x_1 = \log(sales)$, $x_2 = ceoten$, and $x_3 = ceoten^2$. As we know from Chapter 2, the parameter β_1 is the ceteris paribus

TABLE 3.1 Terminology for Multiple Regression

y	$x_1, x_2, \ldots, x_k$
Dependent variable	Independent variables
Explained variable	Explanatory variables
Response variable	Control variables
Predicted variable	Predictor variables
Regressand	Regressors

elasticity of *salary* with respect to *sales*. If $\beta_3 = 0$, then $100\beta_2$ is approximately the ceteris paribus percentage increase in *salary* when *ceoten* increases by one year. When $\beta_3 \neq 0$, the effect of *ceoten* on *salary* is more complicated. We will postpone a detailed treatment of general models with quadratics until Chapter 6.

Equation (3.7) provides an important reminder about multiple regression analysis. The term "linear" in a multiple linear regression model means that equation (3.6) is linear in the *parameters*, β_j. Equation (3.7) is an example of a multiple regression model that, while linear in the β_j, is a nonlinear relationship between *salary* and the variables *sales* and *ceoten*. Many applications of multiple linear regression involve nonlinear relationships among the underlying variables.

The key assumption for the general multiple regression model is easy to state in terms of a conditional expectation:

$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0. \qquad [3.8]$$

At a minimum, equation (3.8) requires that all factors in the unobserved error term be uncorrelated with the explanatory variables. It also means that we have correctly accounted for the functional relationships between the explained and explanatory variables. Any problem that causes u to be correlated with any of the independent variables causes (3.8) to fail. In Section 3-3, we will show that assumption (3.8) implies that OLS is unbiased and will derive the bias that arises when a key variable has been omitted from the equation. In Chapters 15 and 16, we will study other reasons that might cause (3.8) to fail and show what can be done in cases where it does fail.

3-2 Mechanics and Interpretation of Ordinary Least Squares

We now summarize some computational and algebraic features of the method of ordinary least squares as it applies to a particular set of data. We also discuss how to interpret the estimated equation.

3-2a Obtaining the OLS Estimates

We first consider estimating the model with two independent variables. The estimated OLS equation is written in a form similar to the simple regression case:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2, \qquad [3.9]$$

where

$\hat{\beta}_0$ = the estimate of β_0.
$\hat{\beta}_1$ = the estimate of β_1.
$\hat{\beta}_2$ = the estimate of β_2.

But how do we obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? The method of **ordinary least squares** chooses the estimates to minimize the sum of squared residuals. That is, given n observations on y, x_1, and x_2, $\{(x_{i1}, x_{i2}, y_i): i = 1, 2, \ldots, n\}$, the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are chosen simultaneously to make

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \qquad [3.10]$$

as small as possible.

To understand what OLS is doing, it is important to master the meaning of the indexing of the independent variables in (3.10). The independent variables have two subscripts here, i followed by either 1 or 2. The i subscript refers to the observation number. Thus, the sum in (3.10) is over all $i = 1$ to n observations. The second index is simply a method of distinguishing between different independent variables. In the example relating *wage* to *educ* and *exper*, $x_{i1} = educ_i$ is education for person i in the sample and $x_{i2} = exper_i$ is experience for person i. The sum of squared residuals in equation (3.10) is $\sum_{i=1}^{n} (wage_i - \hat{\beta}_0 - \hat{\beta}_1 educ_i - \hat{\beta}_2 exper_i)^2$. In what follows, the i subscript

is reserved for indexing the observation number. If we write x_{ij}, then this means the i^{th} observation on the j^{th} independent variable. (Some authors prefer to switch the order of the observation number and the variable number, so that x_{1i} is observation i on variable one. But this is just a matter of notational taste.)

In the general case with k independent variables, we seek estimates, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ in the equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k. \qquad [3.11]$$

The OLS estimates, $k + 1$ of them, are chosen to minimize the sum of squared residuals:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2. \qquad [3.12]$$

This minimization problem can be solved using multivariable calculus (see Appendix 3A). This leads to $k + 1$ linear equations in $k + 1$ unknowns $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$:

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) &= 0 \\
\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) &= 0 \\
\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) &= 0 \\
&\vdots \\
\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) &= 0.
\end{aligned} \qquad [3.13]$$

These are often called the OLS **first order conditions**. As with the simple regression model in Section 2-2, the OLS first order conditions can be obtained by the method of moments: under assumption (3.8), $\mathrm{E}(u) = 0$ and $\mathrm{E}(x_j u) = 0$, where $j = 1, 2, \ldots, k$. The equations in (3.13) are the sample counterparts of these population moments, although we have omitted the division by the sample size n.

For even moderately sized n and k, solving the equations in (3.13) by hand calculations is tedious. Nevertheless, modern computers running standard statistics and econometrics software can solve these equations with large n and k very quickly.

There is only one slight caveat: we must assume that the equations in (3.13) can be solved *uniquely* for the $\hat{\beta}_j$. For now, we just assume this, as it is usually the case in well-specified models. In Section 3-3, we state the assumption needed for unique OLS estimates to exist (see Assumption MLR.3).

As in simple regression analysis, equation (3.11) is called the **OLS regression line** or the **sample regression function (SRF)**. We will call $\hat{\beta}_0$ the **OLS intercept estimate** and $\hat{\beta}_1, \ldots, \hat{\beta}_k$ the **OLS slope estimates** (corresponding to the independent variables $x_1, x_2, \ldots, x_k$).

To indicate that an OLS regression has been run, we will either write out equation (3.11) with y and $x_1, \ldots, x_k$ replaced by their variable names (such as *wage*, *educ*, and *exper*), or we will say that "we ran an OLS regression of y on $x_1, x_2, \ldots, x_k$" or that "we regressed y on $x_1, x_2, \ldots, x_k$." These are shorthand for saying that the method of ordinary least squares was used to obtain the OLS equation (3.11). Unless explicitly stated otherwise, we always estimate an intercept along with the slopes.

3-2b Interpreting the OLS Regression Equation

More important than the details underlying the computation of the $\hat{\beta}_j$ is the *interpretation* of the estimated equation. We begin with the case of two independent variables:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \qquad [3.14]$$

The intercept $\hat{\beta}_0$ in equation (3.14) is the predicted value of y when $x_1 = 0$ and $x_2 = 0$. Sometimes, setting x_1 and x_2 both equal to zero is an interesting scenario; in other cases, it will not make sense. Nevertheless, the intercept is always needed to obtain a prediction of y from the OLS regression line, as (3.14) makes clear.

The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ have **partial effect**, or **ceteris paribus**, interpretations. From equation (3.14), we have

$$\Delta\hat{y} = \hat{\beta}_1\Delta x_1 + \hat{\beta}_2\Delta x_2,$$

so we can obtain the predicted change in y given the changes in x_1 and x_2. (Note how the intercept has nothing to do with the changes in y.) In particular, when x_2 is held fixed, so that $\Delta x_2 = 0$, then

$$\Delta\hat{y} = \hat{\beta}_1\Delta x_1,$$

holding x_2 fixed. The key point is that, by including x_2 in our model, we obtain a coefficient on x_1 with a ceteris paribus interpretation. This is why multiple regression analysis is so useful. Similarly,

$$\Delta\hat{y} = \hat{\beta}_2\Delta x_2,$$

holding x_1 fixed.

EXAMPLE 3.1 Determinants of College GPA

The variables in GPA1 include the college grade point average (*colGPA*), high school GPA (*hsGPA*), and achievement test score (*ACT*) for a sample of 141 students from a large university; both college and high school GPAs are on a four-point scale. We obtain the following OLS regression line to predict college GPA from high school GPA and achievement test score:

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT$$
$$n = 141. \qquad [3.15]$$

How do we interpret this equation? First, the intercept 1.29 is the predicted college GPA if *hsGPA* and *ACT* are both set as zero. Because no one who attends college has either a zero high school GPA or a zero on the achievement test, the intercept in this equation is not, by itself, meaningful.

More interesting estimates are the slope coefficients on *hsGPA* and *ACT*. As expected, there is a positive partial relationship between *colGPA* and *hsGPA*: holding *ACT* fixed, another point on *hsGPA* is associated with .453 of a point on the college GPA, or almost half a point. In other words, if we choose two students, A and B, and these students have the same ACT score, but the high school GPA of Student A is one point higher than the high school GPA of Student B, then we predict Student A to have a college GPA .453 higher than that of Student B. (This says nothing about any two actual people, but it is our best prediction.)

The sign on *ACT* implies that, while holding *hsGPA* fixed, a change in the ACT score of 10 points—a very large change, as the maximum ACT score is 36 and the average score in the sample is about 24 with a standard deviation less than three—affects *colGPA* by less than one-tenth of a point. This is a small effect, and it suggests that, once high school GPA is accounted for, the ACT score is not a strong predictor of college GPA. (Naturally, there are many other factors that contribute to GPA, but here we focus on statistics available for high school students.) Later, after we discuss statistical inference, we will show that not only is the coefficient on ACT practically small, it is also statistically insignificant.

If we focus on a simple regression analysis relating *colGPA* to *ACT* only, we obtain

$$\widehat{colGPA} = 2.40 + .0271\, ACT$$
$$n = 141;$$

thus, the coefficient on *ACT* is almost three times as large as the estimate in (3.15). But this equation does *not* allow us to compare two people with the same high school GPA; it corresponds to a different experiment. We say more about the differences between multiple and simple regression later.

The case with more than two independent variables is similar. The OLS regression line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k. \quad [3.16]$$

Written in terms of changes,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \cdots + \hat{\beta}_k \Delta x_k. \quad [3.17]$$

The coefficient on x_1 measures the change in $\hat{y}$ due to a one-unit increase in x_1, holding all other independent variables fixed. That is,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1, \quad [3.18]$$

holding $x_2, x_3, \ldots, x_k$ fixed. Thus, we have *controlled for* the variables $x_2, x_3, \ldots, x_k$ when estimating the effect of x_1 on y. The other coefficients have a similar interpretation.

The following is an example with three independent variables.

EXAMPLE 3.2 **Hourly Wage Equation**

Using the 526 observations on workers in WAGE1, we include *educ* (years of education), *exper* (years of labor market experience), and *tenure* (years with the current employer) in an equation explaining log(*wage*). The estimated equation is

$$\widehat{\log(wage)} = .284 + .092\, educ + .0041\, exper + .022\, tenure$$
$$n = 526. \quad [3.19]$$

As in the simple regression case, the coefficients have a percentage interpretation. The only difference here is that they also have a ceteris paribus interpretation. The coefficient .092 means that, holding *exper* and *tenure* fixed, another year of education is predicted to increase log(*wage*) by .092, which translates into an approximate 9.2% [100(.092)] increase in *wage*. Alternatively, if we take two people with the same levels of experience and job tenure, the coefficient on *educ* is the proportionate difference in predicted wage when their education levels differ by one year. This measure of the return to education at least keeps two important productivity factors fixed; whether it is a good estimate of the ceteris paribus return to another year of education requires us to study the statistical properties of OLS (see Section 3-3).

3-2c On the Meaning of "Holding Other Factors Fixed" in Multiple Regression

The partial effect interpretation of slope coefficients in multiple regression analysis can cause some confusion, so we provide a further discussion now.

In Example 3.1, we observed that the coefficient on *ACT* measures the predicted difference in *colGPA*, holding *hsGPA* fixed. The power of multiple regression analysis is that it provides this ceteris paribus interpretation even though the data have *not* been collected in a ceteris paribus fashion. In giving the coefficient on *ACT* a partial effect interpretation, it may seem that we actually went out and sampled people with the same high school GPA but possibly with different ACT scores. This is not the case. The data are a random sample from a large university: there were no restrictions placed on the sample values of *hsGPA* or *ACT* in obtaining the data. Rarely do we have the luxury of holding certain variables fixed in obtaining our sample. *If* we could collect a sample of individuals with the same high school GPA, then we could perform a simple regression analysis relating *colGPA* to *ACT*. Multiple regression effectively allows us to mimic this situation without restricting the values of any independent variables.

The power of multiple regression analysis is that it allows us to do in nonexperimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed.

3-2d Changing More Than One Independent Variable Simultaneously

Sometimes, we want to change more than one independent variable at the same time to find the resulting effect on the dependent variable. This is easily done using equation (3.17). For example, in equation (3.19), we can obtain the estimated effect on *wage* when an individual stays at the same firm for another year: *exper* (general workforce experience) and *tenure* both increase by one year. The total effect (holding *educ* fixed) is

$$\widehat{\Delta \log(wage)} = .0041\,\Delta exper + .022\,\Delta tenure = .0041 + .022 = .0261,$$

or about 2.6%. Because *exper* and *tenure* each increase by one year, we just add the coefficients on *exper* and *tenure* and multiply by 100 to turn the effect into a percentage.

3-2e OLS Fitted Values and Residuals

After obtaining the OLS regression line (3.11), we can obtain a *fitted* or *predicted value* for each observation. For observation i, the fitted value is simply

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}, \qquad [3.20]$$

which is just the predicted value obtained by plugging the values of the independent variables for observation i into equation (3.11). We should not forget about the intercept in obtaining the fitted values; otherwise, the answer can be very misleading. As an example, if in (3.15), $hsGPA_i = 3.5$ and $ACT_i = 24$, $\widehat{colGPA}_i = 1.29 + .453(3.5) + .0094(24) = 3.101$ (rounded to three places after the decimal).

Normally, the actual value y_i for any observation i will not equal the predicted value, $\hat{y}_i$: OLS minimizes the *average* squared prediction error, which says nothing about the prediction error for any particular observation. The **residual** for observation i is defined just as in the simple regression case,

$$\hat{u}_i = y_i - \hat{y}_i. \qquad [3.21]$$

There is a residual for each observation. If $\hat{u}_i > 0$, then $\hat{y}_i$ is below y_i, which means that, for this observation, y_i is underpredicted. If $\hat{u}_i < 0$, then $y_i < \hat{y}_i$, and y_i is overpredicted.

The OLS fitted values and residuals have some important properties that are immediate extensions from the single variable case:

1. The sample average of the residuals is zero and so $\bar{y} = \bar{\hat{y}}$.
2. The sample covariance between each independent variable and the OLS residuals is zero. Consequently, the sample covariance between the OLS fitted values and the OLS residuals is zero.
3. The point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$ is always on the OLS regression line: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \cdots + \hat{\beta}_k\bar{x}_k$.

The first two properties are immediate consequences of the set of equations used to obtain the OLS estimates. The first equation in (3.13) says that the sum of the residuals is zero. The remaining equations are of the form $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, which implies that each independent variable has zero sample covariance with $\hat{u}_i$. Property (3) follows immediately from property (1).

GOING FURTHER 3.2

In Example 3.1, the OLS fitted line explaining college GPA in terms of high school GPA and ACT score is

$$\widehat{colGPA} = 1.29 + .453\,hsGPA + .0094\,ACT.$$

If the average high school GPA is about 3.4 and the average ACT score is about 24.2, what is the average college GPA in the sample?

3-2f A "Partialling Out" Interpretation of Multiple Regression

When applying OLS, we do not need to know explicit formulas for the $\hat{\beta}_j$ that solve the system of equations in (3.13). Nevertheless, for certain derivations, we do need explicit formulas for the $\hat{\beta}_j$. These formulas also shed further light on the workings of OLS.

Consider again the case with $k = 2$ independent variables, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. For concreteness, we focus on $\hat{\beta}_1$. One way to express $\hat{\beta}_1$ is

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} \hat{r}_{i1} y_i\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right), \qquad [3.22]$$

where the $\hat{r}_{i1}$ are the OLS residuals from a simple regression of x_1 on x_2, using the sample at hand. We regress our first independent variable, x_1, on our second independent variable, x_2, and then obtain the residuals (y plays no role here). Equation (3.22) shows that we can then do a simple regression of y on $\hat{r}_1$ to obtain $\hat{\beta}_1$. (Note that the residuals $\hat{r}_{i1}$ have a zero sample average, and so $\hat{\beta}_1$ is the usual slope estimate from simple regression.)

The representation in equation (3.22) gives another demonstration of $\hat{\beta}_1$'s partial effect interpretation. The residuals $\hat{r}_{i1}$ are the part of x_{i1} that is uncorrelated with x_{i2}. Another way of saying this is that $\hat{r}_{i1}$ is x_{i1} after the effects of x_{i2} have been *partialled out*, or *netted out*. Thus, $\hat{\beta}_1$ measures the sample relationship between y and x_1 after x_2 has been partialled out.

In simple regression analysis, there is no partialling out of other variables because no other variables are included in the regression. Computer Exercise C5 steps you through the partialling out process using the wage data from Example 3.2. For practical purposes, the important thing is that $\hat{\beta}_1$ in the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ measures the change in y given a one-unit increase in x_1, holding x_2 fixed.

In the general model with k explanatory variables, $\hat{\beta}_1$ can still be written as in equation (3.22), but the residuals $\hat{r}_{i1}$ come from the regression of x_1 on $x_2, \ldots, x_k$. Thus, $\hat{\beta}_1$ measures the effect of x_1 on y after $x_2, \ldots, x_k$ have been partialled or netted out. In econometrics, the general partialling out result is usually called the **Frisch-Waugh theorem**. It has many uses in theoretical and applied econometrics. We will see applications to time series regressions in Chapter 10.

3-2g Comparison of Simple and Multiple Regression Estimates

Two special cases exist in which the simple regression of y on x_1 will produce the *same* OLS estimate on x_1 as the regression of y on x_1 and x_2. To be more precise, write the simple regression of y on x_1 as $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$, and write the multiple regression as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. We know that the simple regression coefficient $\tilde{\beta}_1$ does not usually equal the multiple regression coefficient $\hat{\beta}_1$. It turns out there is a simple relationship between $\tilde{\beta}_1$ and $\hat{\beta}_1$, which allows for interesting comparisons between simple and multiple regression:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1, \qquad [3.23]$$

where $\tilde{\delta}_1$ is the slope coefficient from the simple regression of x_{i2} on x_{i1}, $i = 1, \ldots, n$. This equation shows how $\tilde{\beta}_1$ differs from the partial effect of x_1 on $\hat{y}$. The confounding term is the partial effect of x_2 on $\hat{y}$ times the slope in the simple regression of x_2 on x_1. (See Section 3A-4 in the chapter appendix for a more general verification.)

The relationship between $\tilde{\beta}_1$ and $\hat{\beta}_1$ also shows there are two distinct cases where they are equal:

1. The partial effect of x_2 on $\hat{y}$ is zero in the sample. That is, $\hat{\beta}_2 = 0$.
2. x_1 and x_2 are uncorrelated in the sample. That is, $\tilde{\delta}_1 = 0$.

Even though simple and multiple regression estimates are almost never identical, we can use the above formula to characterize why they might be either very different or quite similar. For example, if $\hat{\beta}_2$ is small, we might expect the multiple and simple regression estimates of β_1 to be similar.

In Example 3.1, the sample correlation between *hsGPA* and *ACT* is about .346, which is a nontrivial correlation. But the coefficient on *ACT* is pretty small. Therefore, it is not surprising to find that the simple regression of *colGPA* on *hsGPA* produces a slope estimate of .482, which is not much different from the estimate .453 in (3.15).

EXAMPLE 3.3 **Participation in 401(k) Pension Plans**

We use the data in 401K to estimate the effect of a plan's match rate (*mrate*) on the participation rate (*prate*) in its 401(k) pension plan. The match rate is the amount the firm contributes to a worker's fund for each dollar the worker contributes (up to some limit); thus, $mrate = .75$ means that the firm contributes 75¢ for each dollar contributed by the worker. The participation rate is the percentage of eligible workers having a 401(k) account. The variable *age* is the age of the 401(k) plan. There are 1,534 plans in the data set, the average *prate* is 87.36, the average *mrate* is .732, and the average *age* is 13.2.

Regressing *prate* on *mrate*, *age* gives

$$\widehat{prate} = 80.12 + 5.52\, mrate + .243\, age$$

$$n = 1{,}534.$$

Thus, both *mrate* and *age* have the expected effects. What happens if we do not control for *age*? The estimated effect of *age* is not trivial, and so we might expect a large change in the estimated effect of *mrate* if *age* is dropped from the regression. However, the simple regression of *prate* on *mrate* yields $\widehat{prate} = 83.08 + 5.86\, mrate$. The simple regression estimate of the effect of *mrate* on *prate* is clearly different from the multiple regression estimate, but the difference is not very big. (The simple regression estimate is only about 6.2% larger than the multiple regression estimate.) This can be explained by the fact that the sample correlation between *mrate* and *age* is only .12.

In the case with k independent variables, the simple regression of y on x_1 and the multiple regression of y on $x_1, x_2, \ldots, x_k$ produce an identical estimate of x_1 only if (1) the OLS coefficients on x_2 through x_k are all zero *or* (2) x_1 is uncorrelated with *each of* $x_2, \ldots, x_k$. Neither of these is very likely in practice. But if the coefficients on x_2 through x_k are small, or the sample correlations between x_1 and the other independent variables are insubstantial, then the simple and multiple regression estimates of the effect of x_1 on y can be similar.

3-2h Goodness-of-Fit

As with simple regression, we can define the **total sum of squares (SST)**, the **explained sum of squares (SSE)**, and the **residual sum of squares** or **sum of squared residuals (SSR)** as

$$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad [3.24]$$

$$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad [3.25]$$

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2. \qquad [3.26]$$

Using the same argument as in the simple regression case, we can show that

$$SST = SSE + SSR. \qquad [3.27]$$

In other words, the total variation in $\{y_i\}$ is the sum of the total variations in $\{\hat{y}_i\}$ and in $\{\hat{u}_i\}$.

Assuming that the total variation in y is nonzero, as is the case unless y_i is constant in the sample, we can divide (3.27) by SST to get

$$\text{SSR/SST} + \text{SSE/SST} = 1.$$

Just as in the simple regression case, the R-squared is defined to be

$$R^2 \equiv \text{SSE/SST} = 1 - \text{SSR/SST}, \qquad [3.28]$$

and it is interpreted as the proportion of the sample variation in y_i that is explained by the OLS regression line. By definition, R^2 is a number between zero and one.

R^2 can also be shown to equal the squared correlation coefficient between the actual y_i and the fitted values $\hat{y}_i$. That is,

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)}. \qquad [3.29]$$

[We have put the average of the $\hat{y}_i$ in (3.29) to be true to the formula for a correlation coefficient; we know that this average equals $\bar{y}$ because the sample average of the residuals is zero and $y_i = \hat{y}_i + \hat{u}_i$.]

EXAMPLE 3.4 **Determinants of College GPA**

From the grade point average regression that we did earlier, the equation with R^2 is

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT$$

$$n = 141,\ R^2 = .176.$$

This means that *hsGPA* and *ACT* together explain about 17.6% of the variation in college GPA for this sample of students. This may not seem like a high percentage, but we must remember that there are many other factors—including family background, personality, quality of high school education, affinity for college—that contribute to a student's college performance. If *hsGPA* and *ACT* explained almost all of the variation in *colGPA*, then performance in college would be preordained by high school performance!

An important fact about R^2 is that it never decreases, and it usually increases, when another independent variable is added to a regression and the same set of observations is used for both regressions. This algebraic fact follows because, by definition, the sum of squared residuals never increases when additional regressors are added to the model. For example, the last digit of one's social security number has nothing to do with one's hourly wage, but adding this digit to a wage equation will increase the R^2 (by a little, at least).

An important caveat to the previous assertion about R-squared is that it assumes we do not have missing data on the explanatory variables. If two regressions use different sets of observations, then, in general, we cannot tell how the R-squareds will compare, even if one regression uses a subset of regressors. For example, suppose we have a full set of data on the variables y, x_1, and x_2, but for some units in our sample data are missing on x_3. Then we cannot say that the R-squared from regressing y on x_1, x_2 will be less than that from regressing y on x_1, x_2, and x_3: it could go either way. Missing data can be an important practical issue, and we will return to it in Chapter 9.

The fact that R^2 never decreases when *any* variable is added to a regression makes it a poor tool for deciding whether one variable or several variables should be added to a model. The factor that should determine whether an explanatory variable belongs in a model is whether the explanatory

variable has a nonzero partial effect on y in the *population*. We will show how to test this hypothesis in Chapter 4 when we cover statistical inference. We will also see that, when used properly, R^2 allows us to *test* a group of variables to see if it is important for explaining y. For now, we use it as a goodness-of-fit measure for a given model.

EXAMPLE 3.5 Explaining Arrest Records

CRIME1 contains data on arrests during the year 1986 and other information on 2,725 men born in either 1960 or 1961 in California. Each man in the sample was arrested at least once prior to 1986. The variable *narr86* is the number of times the man was arrested during 1986: it is zero for most men in the sample (72.29%), and it varies from 0 to 12. (The percentage of men arrested once during 1986 was 20.51.) The variable *pcnv* is the proportion (not percentage) of arrests prior to 1986 that led to conviction, *avgsen* is average sentence length served for prior convictions (zero for most people), *ptime86* is months spent in prison in 1986, and *qemp86* is the number of quarters during which the man was employed in 1986 (from zero to four).

A linear model explaining arrests is

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 ptime86 + \beta_4 qemp86 + u,$$

where *pcnv* is a proxy for the likelihood for being convicted of a crime and *avgsen* is a measure of expected severity of punishment, if convicted. The variable *ptime86* captures the incarcerative effects of crime: if an individual is in prison, he cannot be arrested for a crime outside of prison. Labor market opportunities are crudely captured by *qemp86*.

First, we estimate the model without the variable *avgsen*. We obtain

$$\widehat{narr86} = .712 - .150\, pcnv - .034\, ptime86 - .104\, qemp86$$

$$n = 2{,}725,\ R^2 = .0413.$$

This equation says that, as a group, the three variables *pcnv*, *ptime86*, and *qemp86* explain about 4.1% of the variation in *narr86*.

Each of the OLS slope coefficients has the anticipated sign. An increase in the proportion of convictions lowers the predicted number of arrests. If we increase *pcnv* by .50 (a large increase in the probability of conviction), then, holding the other factors fixed, $\Delta\widehat{narr86} = -.150(.50) = -.075$. This may seem unusual because an arrest cannot change by a fraction. But we can use this value to obtain the predicted change in expected arrests for a large group of men. For example, among 100 men, the predicted fall in arrests when *pcnv* increases by .50 is -7.5.

Similarly, a longer prison term leads to a lower predicted number of arrests. In fact, if *ptime86* increases from 0 to 12, predicted arrests for a particular man fall by $.034(12) = .408$. Another quarter in which legal employment is reported lowers predicted arrests by .104, which would be 10.4 arrests among 100 men.

If *avgsen* is added to the model, we know that R^2 will increase. The estimated equation is

$$\widehat{narr86} = .707 - .151\, pcnv + .0074\, avgsen - .037\, ptime86 - .103\, qemp86$$

$$n = 2{,}725,\ R^2 = .0422.$$

Thus, adding the average sentence variable increases R^2 from .0413 to .0422, a practically small effect. The sign of the coefficient on *avgsen* is also unexpected: it says that a longer average sentence length increases criminal activity.

Example 3.5 deserves a final word of caution. The fact that the four explanatory variables included in the second regression explain only about 4.2% of the variation in *narr86* does not necessarily

mean that the equation is useless. Even though these variables collectively do not explain much of the variation in arrests, it is still possible that the OLS estimates are reliable estimates of the ceteris paribus effects of each independent variable on *narr86*. As we will see, whether this is the case does not directly depend on the size of R^2. Generally, a low R^2 indicates that it is hard to predict individual outcomes on y with much accuracy, something we study in more detail in Chapter 6. In the arrest example, the small R^2 reflects what we already suspect in the social sciences: it is generally very difficult to predict individual behavior.

3-2i Regression through the Origin

Sometimes, an economic theory or common sense suggests that β_0 should be zero, and so we should briefly mention OLS estimation when the intercept is zero. Specifically, we now seek an equation of the form

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \cdots + \tilde{\beta}_k x_k, \qquad [3.30]$$

where the symbol "~" over the estimates is used to distinguish them from the OLS estimates obtained along with the intercept [as in (3.11)]. In (3.30), when $x_1 = 0, x_2 = 0, \ldots, x_k = 0$, the predicted value is zero. In this case, $\tilde{\beta}_1, \ldots, \tilde{\beta}_k$ are said to be the OLS estimates from the regression of y on $x_1, x_2, \ldots, x_k$ *through the origin*.

The OLS estimates in (3.30), as always, minimize the sum of squared residuals, but with the intercept set at zero. You should be warned that the properties of OLS that we derived earlier no longer hold for regression through the origin. In particular, the OLS residuals no longer have a zero sample average. Further, if R^2 is defined as $1 - \text{SSR/SST}$, where SST is given in (3.24) and SSR is now $\sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_k x_{ik})^2$, then R^2 can actually be negative. This means that the sample average, $\bar{y}$, "explains" more of the variation in the y_i than the explanatory variables. Either we should include an intercept in the regression or conclude that the explanatory variables poorly explain y. To always have a nonnegative R-squared, some economists prefer to calculate R^2 as the squared correlation coefficient between the actual and fitted values of y, as in (3.29). (In this case, the average fitted value must be computed directly because it no longer equals $\bar{y}$.) However, there is no set rule on computing R-squared for regression through the origin.

One serious drawback with regression through the origin is that, if the intercept β_0 in the population model is different from zero, then the OLS estimators of the slope parameters will be biased. The bias can be severe in some cases. The cost of estimating an intercept when β_0 is truly zero is that the variances of the OLS slope estimators are larger.

3-3 The Expected Value of the OLS Estimators

We now turn to the statistical properties of OLS for estimating the parameters in an underlying population model. In this section, we derive the expected value of the OLS estimators. In particular, we state and discuss four assumptions, which are direct extensions of the simple regression model assumptions, under which the OLS estimators are unbiased for the population parameters. We also explicitly obtain the bias in OLS when an important variable has been omitted from the regression.

You should remember that statistical properties have nothing to do with a particular sample, but rather with the property of estimators when random sampling is done repeatedly. Thus, Sections 3-3, 3-4, and 3-5 are somewhat abstract. Although we give examples of deriving bias for particular models, it is not meaningful to talk about the statistical properties of a set of estimates obtained from a single sample.

The first assumption we make simply defines the multiple linear regression (MLR) model.

Assumption MLR.1 Linear in Parameters

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u, \qquad [3.31]$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term.

Equation (3.31) formally states the **population model**, sometimes called the **true model**, to allow for the possibility that we might estimate a model that differs from (3.31). The key feature is that the model is linear in the parameters $\beta_0, \beta_1, \ldots, \beta_k$. As we know, (3.31) is quite flexible because y and the independent variables can be arbitrary functions of the underlying variables of interest, such as natural logarithms and squares [see, for example, equation (3.7)].

Assumption MLR.2 Random Sampling

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

Sometimes, we need to write the equation for a particular observation i: for a randomly drawn observation from the population, we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i. \qquad [3.32]$$

Remember that i refers to the observation, and the second subscript on x is the variable number. For example, we can write a CEO salary equation for a particular CEO i as

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ceoten_i + \beta_3 ceoten_i^2 + u_i. \qquad [3.33]$$

The term u_i contains the unobserved factors for CEO i that affect his or her salary. For applications, it is usually easiest to write the model in population form, as in (3.31). It contains less clutter and emphasizes the fact that we are interested in estimating a population relationship.

In light of model (3.31), the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ from the regression of y on $x_1, \ldots, x_k$ are now considered to be estimators of $\beta_0, \beta_1, \ldots, \beta_k$. In Section 3-2, we saw that OLS chooses the intercept and slope estimates for a particular sample so that the residuals average to zero and the sample correlation between each independent variable and the residuals is zero. Still, we did not include conditions under which the OLS estimates are well defined for a given sample. The next assumption fills that gap.

Assumption MLR.3 No Perfect Collinearity

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

Assumption MLR.3 is more complicated than its counterpart for simple regression because we must now look at relationships between all independent variables. If an independent variable in (3.31) is an exact linear combination of the other independent variables, then we say the model suffers from **perfect collinearity**, and it cannot be estimated by OLS.

It is important to note that Assumption MLR.3 *does* allow the independent variables to be correlated; they just cannot be *perfectly* correlated. If we did not allow for any correlation among the

independent variables, then multiple regression would be of very limited use for econometric analysis. For example, in the model relating test scores to educational expenditures and average family income,

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u,$$

we fully expect *expend* and *avginc* to be correlated: school districts with high average family incomes tend to spend more per student on education. In fact, the primary motivation for including *avginc* in the equation is that we suspect it is correlated with *expend*, and so we would like to hold it fixed in the analysis. Assumption MLR.3 only rules out *perfect* correlation between *expend* and *avginc* in our sample. We would be very unlucky to obtain a sample where per-student expenditures are perfectly correlated with average family income. But some correlation, perhaps a substantial amount, is expected and certainly allowed.

The simplest way that two independent variables can be perfectly correlated is when one variable is a constant multiple of another. This can happen when a researcher inadvertently puts the same variable measured in different units into a regression equation. For example, in estimating a relationship between consumption and income, it makes no sense to include as independent variables income measured in dollars as well as income measured in thousands of dollars. One of these is redundant. What sense would it make to hold income measured in dollars fixed while changing income measured in thousands of dollars?

We already know that different nonlinear functions of the same variable *can* appear among the regressors. For example, the model $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ does not violate Assumption MLR.3: even though $x_2 = inc^2$ is an exact function of $x_1 = inc$, inc^2 is not an exact *linear* function of *inc*. Including inc^2 in the model is a useful way to generalize functional form, unlike including income measured in dollars and in thousands of dollars.

Common sense tells us not to include the same explanatory variable measured in different units in the same regression equation. There are also more subtle ways that one independent variable can be a multiple of another. Suppose we would like to estimate an extension of a constant elasticity consumption function. It might seem natural to specify a model such as

$$\log(cons) = \beta_0 + \beta_1 \log(inc) + \beta_2 \log(inc^2) + u, \qquad [3.34]$$

where $x_1 = \log(inc)$ and $x_2 = \log(inc^2)$. Using the basic properties of the natural log (see Math Refresher A), $\log(inc^2) = 2 \cdot \log(inc)$. That is, $x_2 = 2x_1$, and naturally this holds for all observations in the sample. This violates Assumption MLR.3. What we should do instead is include $[\log(inc)]^2$, not $\log(inc^2)$, along with $\log(inc)$. This is a sensible extension of the constant elasticity model, and we will see how to interpret such models in Chapter 6.

Another way that independent variables can be perfectly collinear is when one independent variable can be expressed as an exact linear function of two or more of the other independent variables. For example, suppose we want to estimate the effect of campaign spending on campaign outcomes. For simplicity, assume that each election has two candidates. Let *voteA* be the percentage of the vote for Candidate A, let *expendA* be campaign expenditures by Candidate A, let *expendB* be campaign expenditures by Candidate B, and let *totexpend* be total campaign expenditures; the latter three variables are all measured in dollars. It may seem natural to specify the model as

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend + u, \qquad [3.35]$$

in order to isolate the effects of spending by each candidate and the total amount of spending. But this model violates Assumption MLR.3 because $x_3 = x_1 + x_2$ by definition. Trying to interpret this equation in a ceteris paribus fashion reveals the problem. The parameter of β_1 in equation (3.35) is supposed to measure the effect of increasing expenditures by Candidate A by one dollar on Candidate A's vote, holding Candidate B's spending *and* total spending fixed. This is nonsense, because if *expendB* and *totexpend* are held fixed, then we cannot increase *expendA*.

The solution to the perfect collinearity in (3.35) is simple: drop any one of the three variables from the model. We would probably drop *totexpend*, and then the coefficient on *expendA* would

measure the effect of increasing expenditures by A on the percentage of the vote received by A, holding the spending by B fixed.

The prior examples show that Assumption MLR.3 can fail if we are not careful in specifying our model. Assumption MLR.3 also fails if the sample size, n, is too small in relation to the number of parameters being estimated. In the general regression model in equation (3.31), there are $k + 1$ parameters, and MLR.3 fails if $n < k + 1$. Intuitively, this makes sense: to estimate $k + 1$ parameters, we need at least $k + 1$ observations. Not surprisingly, it is better to have as many observations as possible, something we will see with our variance calculations in Section 3-4.

GOING FURTHER 3.3

In the previous example, if we use as explanatory variables *expendA*, *expendB*, and *shareA*, where *shareA* = 100·(*expendA*/*totexpend*) is the percentage share of total campaign expenditures made by Candidate A, does this violate Assumption MLR.3?

If the model is carefully specified and $n \geq k + 1$, Assumption MLR.3 can fail in rare cases due to bad luck in collecting the sample. For example, in a wage equation with education and experience as variables, it is possible that we could obtain a random sample where each individual has exactly twice as much education as years of experience. This scenario would cause Assumption MLR.3 to fail, but it can be considered very unlikely unless we have an extremely small sample size.

The final, and most important, assumption needed for unbiasedness is a direct extension of Assumption SLR.4.

Assumption MLR.4 Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0. \qquad [3.36]$$

One way that Assumption MLR.4 can fail is if the functional relationship between the explained and explanatory variables is misspecified in equation (3.31): for example, if we forget to include the quadratic term inc^2 in the consumption function $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ when we estimate the model. Another functional form misspecification occurs when we use the level of a variable when the log of the variable is what actually shows up in the population model, or vice versa. For example, if the true model has log(*wage*) as the dependent variable but we use *wage* as the dependent variable in our regression analysis, then the estimators will be biased. Intuitively, this should be pretty clear. We will discuss ways of detecting functional form misspecification in Chapter 9.

Omitting an important factor that is correlated with any of $x_1, x_2, \ldots, x_k$ causes Assumption MLR.4 to fail also. With multiple regression analysis, we are able to include many factors among the explanatory variables, and omitted variables are less likely to be a problem in multiple regression analysis than in simple regression analysis. Nevertheless, in any application, there are always factors that, due to data limitations or ignorance, we will not be able to include. If we think these factors should be controlled for and they are correlated with one or more of the independent variables, then Assumption MLR.4 will be violated. We will derive this bias later.

There are other ways that u can be correlated with an explanatory variable. In Chapters 9 and 15, we will discuss the problem of measurement error in an explanatory variable. In Chapter 16, we cover the conceptually more difficult problem in which one or more of the explanatory variables is determined jointly with y—as occurs when we view quantities and prices as being determined by the intersection of supply and demand curves. We must postpone our study of these problems until we have a firm grasp of multiple regression analysis under an ideal set of assumptions.

When Assumption MLR.4 holds, we often say that we have **exogenous explanatory variables**. If x_j is correlated with u for any reason, then x_j is said to be an **endogenous explanatory variable**.

The terms "exogenous" and "endogenous" originated in simultaneous equations analysis (see Chapter 16), but the term "endogenous explanatory variable" has evolved to cover any case in which an explanatory variable may be correlated with the error term.

Before we show the unbiasedness of the OLS estimators under MLR.1 to MLR.4, a word of caution. Beginning students of econometrics sometimes confuse Assumptions MLR.3 and MLR.4, but they are quite different. Assumption MLR.3 rules out certain relationships among the independent or explanatory variables and has *nothing* to do with the error, u. You will know immediately when carrying out OLS estimation whether or not Assumption MLR.3 holds. On the other hand, Assumption MLR.4—the much more important of the two—restricts the relationship between the unobserved factors in u and the explanatory variables. Unfortunately, we will never know for sure whether the average value of the unobserved factors is unrelated to the explanatory variables. But this is the critical assumption.

We are now ready to show unbiasedness of OLS under the first four multiple regression assumptions. As in the simple regression case, the expectations are conditional on the values of the explanatory variables in the sample, something we show explicitly in Appendix 3A but not in the text.

THEOREM 3.1

UNBIASEDNESS OF OLS

Under Assumptions MLR.1 through MLR.4,

$$\mathrm{E}(\hat{\beta}_j) = \beta_j, \; j = 0, 1, \ldots, k, \qquad [3.37]$$

for any values of the population parameter β_j. In other words, the OLS estimators are unbiased estimators of the population parameters.

In our previous empirical examples, Assumption MLR.3 has been satisfied (because we have been able to compute the OLS estimates). Furthermore, for the most part, the samples are randomly chosen from a well-defined population. If we believe that the specified models are correct under the key Assumption MLR.4, then we can conclude that OLS is unbiased in these examples.

Because we are approaching the point where we can use multiple regression in serious empirical work, it is useful to remember the meaning of unbiasedness. It is tempting, in examples such as the wage equation in (3.19), to say something like "9.2% is an unbiased estimate of the return to education." As we know, an estimate cannot be unbiased: an estimate is a fixed number, obtained from a particular sample, which usually is not equal to the population parameter. When we say that OLS is unbiased under Assumptions MLR.1 through MLR.4, we mean that the *procedure* by which the OLS estimates are obtained is unbiased when we view the procedure as being applied across all possible random samples. We hope that we have obtained a sample that gives us an estimate close to the population value, but, unfortunately, this cannot be assured. What is assured is that we have no reason to believe our estimate is more likely to be too big or more likely to be too small.

3-3a Including Irrelevant Variables in a Regression Model

One issue that we can dispense with fairly quickly is that of **inclusion of an irrelevant variable** or **overspecifying the model** in multiple regression analysis. This means that one (or more) of the independent variables is included in the model even though it has no partial effect on y in the population. (That is, its population coefficient is zero.)

To illustrate the issue, suppose we specify the model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u, \qquad [3.38]$$

and this model satisfies Assumptions MLR.1 through MLR.4. However, x_3 has no effect on y after x_1 and x_2 have been controlled for, which means that $\beta_3 = 0$. The variable x_3 may or may not be

correlated with x_1 or x_2; all that matters is that, once x_1 and x_2 are controlled for, x_3 has no effect on y. In terms of conditional expectations, $E(y|x_1, x_2, x_3) = E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Because we do not know that $\beta_3 = 0$, we are inclined to estimate the equation including x_3:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3. \qquad [3.39]$$

We have included the irrelevant variable, x_3, in our regression. What is the effect of including x_3 in (3.39) when its coefficient in the population model (3.38) is zero? In terms of the unbiasedness of $\hat{\beta}_1$ and $\hat{\beta}_2$, there is *no effect*. This conclusion requires no special derivation, as it follows immediately from Theorem 3.1. Remember, unbiasedness means $E(\hat{\beta}_j) = \beta_j$ for *any* value of β_j, including $\beta_j = 0$. Thus, we can conclude that $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_2) = \beta_2$, $E(\hat{\beta}_3) = 0$ (for any values of β_0, β_1, and β_2). Even though $\hat{\beta}_3$ itself will never be exactly zero, its average value across all random samples will be zero.

The conclusion of the preceding example is much more general: including one or more irrelevant variables in a multiple regression model, or overspecifying the model, does not affect the unbiasedness of the OLS estimators. Does this mean it is harmless to include irrelevant variables? No. As we will see in Section 3-4, including irrelevant variables can have undesirable effects on the *variances* of the OLS estimators.

3-3b Omitted Variable Bias: The Simple Case

Now suppose that, rather than including an irrelevant variable, we omit a variable that actually belongs in the true (or population) model. This is often called the problem of **excluding a relevant variable** or **underspecifying the model**. We claimed in Chapter 2 and earlier in this chapter that this problem generally causes the OLS estimators to be biased. It is time to show this explicitly and, just as importantly, to derive the direction and size of the bias.

Deriving the bias caused by omitting an important variable is an example of **misspecification analysis**. We begin with the case where the true population model has two explanatory variables and an error term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \qquad [3.40]$$

and we assume that this model satisfies Assumptions MLR.1 through MLR.4.

Suppose that our primary interest is in β_1, the partial effect of x_1 on y. For example, y is hourly wage (or log of hourly wage), x_1 is education, and x_2 is a measure of innate ability. In order to get an unbiased estimator of β_1, we *should* run a regression of y on x_1 and x_2 (which gives unbiased estimators of β_0, β_1, and β_2). However, due to our ignorance or data unavailability, we estimate the model by *excluding* x_2. In other words, we perform a simple regression of y on x_1 only, obtaining the equation

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \qquad [3.41]$$

We use the symbol "~" rather than "^" to emphasize that $\tilde{\beta}_1$ comes from an underspecified model.

When first learning about the omitted variable problem, it can be difficult to distinguish between the underlying true model, (3.40) in this case, and the model that we actually estimate, which is captured by the regression in (3.41). It may seem silly to omit the variable x_2 if it belongs in the model, but often we have no choice. For example, suppose that *wage* is determined by

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u. \qquad [3.42]$$

Because ability is not observed, we instead estimate the model

$$wage = \beta_0 + \beta_1 educ + v,$$

where $v = \beta_2 abil + u$. The estimator of β_1 from the simple regression of *wage* on *educ* is what we are calling $\tilde{\beta}_1$.

We derive the expected value of $\tilde{\beta}_1$ conditional on the sample values of x_1 and x_2. Deriving this expectation is not difficult because $\tilde{\beta}_1$ is just the OLS slope estimator from a simple regression, and we have already studied this estimator extensively in Chapter 2. The difference here is that we must analyze its properties when the simple regression model is misspecified due to an omitted variable.

As it turns out, we have done almost all of the work to derive the bias in the simple regression estimator of $\tilde{\beta}_1$. From equation (3.23) we have the algebraic relationship $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the slope estimators (if we could have them) from the multiple regression

$$y_i \text{ on } x_{i1}, x_{i2}, i = 1, \ldots, n \qquad [3.43]$$

and $\tilde{\delta}_1$ is the slope from the simple regression

$$x_{i2} \text{ on } x_{i1}, i = 1, \ldots, n. \qquad [3.44]$$

Because $\tilde{\delta}_1$ depends only on the independent variables in the sample, we treat it as fixed (nonrandom) when computing $\mathrm{E}(\tilde{\beta}_1)$. Further, because the model in (3.40) satisfies Assumptions MLR.1 through MLR.4, we know that $\hat{\beta}_1$ and $\hat{\beta}_2$ would be unbiased for β_1 and β_2, respectively. Therefore,

$$\begin{aligned} \mathrm{E}(\tilde{\beta}_1) &= \mathrm{E}(\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1) = \mathrm{E}(\hat{\beta}_1) + \mathrm{E}(\hat{\beta}_2)\tilde{\delta}_1 \\ &= \beta_1 + \beta_2\tilde{\delta}_1, \end{aligned} \qquad [3.45]$$

which implies the bias in $\tilde{\beta}_1$ is

$$\mathrm{Bias}(\tilde{\beta}_1) = \mathrm{E}(\tilde{\beta}_1) - \beta_1 = \beta_2\tilde{\delta}_1. \qquad [3.46]$$

Because the bias in this case arises from omitting the explanatory variable x_2, the term on the right-hand side of equation (3.46) is often called the **omitted variable bias**.

From equation (3.46), we see that there are two cases where $\tilde{\beta}_1$ is unbiased. The first is pretty obvious: if $\beta_2 = 0$—so that x_2 does not appear in the true model (3.40)—then $\tilde{\beta}_1$ is unbiased. We already know this from the simple regression analysis in Chapter 2. The second case is more interesting. If $\tilde{\delta}_1 = 0$, then $\tilde{\beta}_1$ is unbiased for β_1, even if $\beta_2 \neq 0$.

Because $\tilde{\delta}_1$ is the sample covariance between x_1 and x_2 over the sample variance of x_1, $\tilde{\delta}_1 = 0$ if, and only if, x_1 and x_2 are uncorrelated in the sample. Thus, we have the important conclusion that, if x_1 and x_2 are uncorrelated in the sample, then $\tilde{\beta}_1$ is unbiased. This is not surprising: in Section 3-2, we showed that the simple regression estimator $\tilde{\beta}_1$ and the multiple regression estimator $\tilde{\beta}_1$ are the same when x_1 and x_2 are uncorrelated in the sample. [We can also show that $\tilde{\beta}_1$ is unbiased without conditioning on the x_{i2} if $\mathrm{E}(x_2|x_1) = \mathrm{E}(x_2)$; then, for estimating β_1, leaving x_2 in the error term does not violate the zero conditional mean assumption for the error, once we adjust the intercept.]

When x_1 and x_2 are correlated, $\tilde{\delta}_1$ has the same sign as the correlation between x_1 and x_2: $\tilde{\delta}_1 > 0$ if x_1 and x_2 are positively correlated and $\tilde{\delta}_1 < 0$ if x_1 and x_2 are negatively correlated. The sign of the bias in $\tilde{\beta}_1$ depends on the signs of both β_2 and $\tilde{\delta}_1$ and is summarized in Table 3.2 for the four possible cases when there is bias. Table 3.2 warrants careful study. For example, the bias in $\tilde{\beta}_1$ is positive if $\beta_2 > 0$ (x_2 has a positive effect on y) and x_1 and x_2 are positively correlated, the bias is negative if $\beta_2 > 0$ and x_1 and x_2 are negatively correlated, and so on.

Table 3.2 summarizes the direction of the bias, but the size of the bias is also very important. A small bias of either sign need not be a cause for concern. For example, if the return to education in the population is 8.6% and the bias in the OLS estimator is 0.1% (a tenth of one percentage point), then

TABLE 3.2 Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\mathrm{Corr}(x_1, x_2) > 0$	$\mathrm{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

we would not be very concerned. On the other hand, a bias on the order of three percentage points would be much more serious. The size of the bias is determined by the sizes of β_2 and $\tilde{\delta}_1$.

In practice, because β_2 is an unknown population parameter, we cannot be certain whether β_2 is positive or negative. Nevertheless, we usually have a pretty good idea about the direction of the partial effect of x_2 on y. Further, even though the sign of the correlation between x_1 and x_2 cannot be known if x_2 is not observed, in many cases, we can make an educated guess about whether x_1 and x_2 are positively or negatively correlated.

In the wage equation (3.42), by definition, more ability leads to higher productivity and therefore higher wages: $\beta_2 > 0$. Also, there are reasons to believe that *educ* and *abil* are positively correlated: on average, individuals with more innate ability choose higher levels of education. Thus, the OLS estimates from the simple regression equation $wage = \beta_0 + \beta_1 educ + v$ are *on average* too large. This does not mean that the estimate obtained from our sample is too big. We can only say that if we collect many random samples and obtain the simple regression estimates each time, then the average of these estimates will be greater than β_1.

EXAMPLE 3.6 **Hourly Wage Equation**

Suppose the model $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u$ satisfies Assumptions MLR.1 through MLR.4. The data set in WAGE1 does not contain data on ability, so we estimate β_1 from the simple regression

$$\widehat{\log(wage)} = .584 + .083\, educ$$
$$n = 526,\ R^2 = .186. \qquad [3.47]$$

This is the result from only a single sample, so we cannot say that .083 is greater than β_1; the true return to education could be lower or higher than 8.3% (and we will never know for sure). Nevertheless, we know that the average of the estimates across all random samples would be too large.

As a second example, suppose that, at the elementary school level, the average score for students on a standardized exam is determined by

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 povrate + u, \qquad [3.48]$$

where *expend* is expenditure per student and *povrate* is the poverty rate of the children in the school. Using school district data, we only have observations on the percentage of students with a passing grade and per-student expenditures; we do not have information on poverty rates. Thus, we estimate β_1 from the simple regression of *avgscore* on *expend*.

We can again obtain the likely bias in $\tilde{\beta}_1$. First, β_2 is probably negative: there is ample evidence that children living in poverty score lower, on average, on standardized tests. Second, the average expenditure per student is probably negatively correlated with the poverty rate: the higher the poverty rate, the lower the average per-student spending, so that $\text{Corr}(x_1, x_2) < 0$. From Table 3.2, $\tilde{\beta}_1$ will have a positive bias. This observation has important implications. It could be that the true effect of spending is zero; that is, $\beta_1 = 0$. However, the simple regression estimate of β_1 will usually be greater than zero, and this could lead us to conclude that expenditures are important when they are not.

When reading and performing empirical work in economics, it is important to master the terminology associated with biased estimators. In the context of omitting a variable from model (3.40), if $\text{E}(\tilde{\beta}_1) > \beta_1$, then we say that $\tilde{\beta}_1$ has an **upward bias**. When $\text{E}(\tilde{\beta}_1) < \beta_1$, $\tilde{\beta}_1$ has a **downward bias**. These definitions are the same whether β_1 is positive or negative. The phrase **biased toward zero** refers to cases where $\text{E}(\tilde{\beta}_1)$ is closer to zero than is β_1. Therefore, if β_1 is positive, then $\tilde{\beta}_1$ is biased toward zero if it has a downward bias. On the other hand, if $\beta_1 < 0$, then $\tilde{\beta}_1$ is biased toward zero if it has an upward bias.

3-3c Omitted Variable Bias: More General Cases

Deriving the sign of omitted variable bias when there are multiple regressors in the estimated model is more difficult. We must remember that correlation between a single explanatory variable and the error generally results in *all* OLS estimators being biased. For example, suppose the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \qquad [3.49]$$

satisfies Assumptions MLR.1 through MLR.4. But we omit x_3 and estimate the model as

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2. \qquad [3.50]$$

Now, suppose that x_2 and x_3 are uncorrelated, but that x_1 is correlated with x_3. In other words, x_1 is correlated with the omitted variable, but x_2 is not. It is tempting to think that, while $\tilde{\beta}_1$ is probably biased based on the derivation in the previous subsection, $\tilde{\beta}_2$ is unbiased because x_2 is uncorrelated with x_3. Unfortunately, this is *not* generally the case: both $\tilde{\beta}_1$ and $\tilde{\beta}_2$ will normally be biased. The only exception to this is when x_1 and x_2 are also uncorrelated.

Even in the fairly simple model above, it can be difficult to obtain the direction of bias in $\tilde{\beta}_1$ and $\tilde{\beta}_2$. This is because x_1, x_2, and x_3 can all be pairwise correlated. Nevertheless, an approximation is often practically useful. If we assume that x_1 and x_2 are uncorrelated, then we can study the bias in $\tilde{\beta}_1$ as if x_2 were absent from both the population and the estimated models. In fact, when x_1 and x_2 are uncorrelated, it can be shown that

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i3}}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}.$$

This is just like equation (3.45), but β_3 replaces β_2, and x_3 replaces x_2 in regression (3.44). Therefore, the bias in $\tilde{\beta}_1$ is obtained by replacing β_2 with β_3 and x_2 with x_3 in Table 3.2. If $\beta_3 > 0$ and Corr $(x_1, x_3) > 0$, the bias in $\tilde{\beta}_1$ is positive, and so on.

As an example, suppose we add *exper* to the wage model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u.$$

If *abil* is omitted from the model, the estimators of both β_1 and β_2 are biased, even if we assume *exper* is uncorrelated with *abil*. We are mostly interested in the return to education, so it would be nice if we could conclude that $\tilde{\beta}_1$ has an upward or a downward bias due to omitted ability. This conclusion is not possible without further assumptions. As an *approximation*, let us suppose that, in addition to *exper* and *abil* being uncorrelated, *educ* and *exper* are also uncorrelated. (In reality, they are somewhat negatively correlated.) Because $\beta_3 > 0$ and *educ* and *abil* are positively correlated, $\tilde{\beta}_1$ would have an upward bias, just as if *exper* were not in the model.

The reasoning used in the previous example is often followed as a rough guide for obtaining the likely bias in estimators in more complicated models. Usually, the focus is on the relationship between a particular explanatory variable, say, x_1, and the key omitted factor. Strictly speaking, ignoring all other explanatory variables is a valid practice only when each one is uncorrelated with x_1, but it is still a useful guide. Appendix 3A contains a more careful analysis of omitted variable bias with multiple explanatory variables.

3-4 The Variance of the OLS Estimators

We now obtain the variance of the OLS estimators so that, in addition to knowing the central tendencies of the $\hat{\beta}_j$, we also have a measure of the spread in its sampling distribution. Before finding the variances, we add a homoskedasticity assumption, as in Chapter 2. We do this for two reasons.

First, the formulas are simplified by imposing the constant error variance assumption. Second, in Section 3-5, we will see that OLS has an important efficiency property if we add the homoskedasticity assumption.

In the multiple regression framework, homoskedasticity is stated as follows:

Assumption MLR.5 Homoskedasticity

The error u has the same variance given any value of the explanatory variables. In other words, $\text{Var}(u|x_1, \ldots, x_k) = \sigma^2$.

Assumption MLR.5 means that the variance in the error term, u, conditional on the explanatory variables, is the *same* for all combinations of outcomes of the explanatory variables. If this assumption fails, then the model exhibits heteroskedasticity, just as in the two-variable case.

In the equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u,$$

homoskedasticity requires that the variance of the unobserved error u does not depend on the levels of education, experience, or tenure. That is,

$$\text{Var}(u|educ, exper, tenure) = \sigma^2.$$

If this variance changes with any of the three explanatory variables, then heteroskedasticity is present.

Assumptions MLR.1 through MLR.5 are collectively known as the **Gauss-Markov assumptions** (for cross-sectional regression). So far, our statements of the assumptions are suitable only when applied to cross-sectional analysis with random sampling. As we will see, the Gauss-Markov assumptions for time series analysis, and for other situations such as panel data analysis, are more difficult to state, although there are many similarities.

In the discussion that follows, we will use the symbol $\mathbf{x}$ to denote the set of all independent variables, $(x_1, \ldots, x_k)$. Thus, in the wage regression with *educ*, *exper*, and *tenure* as independent variables, $\mathbf{x} = (educ, exper, tenure)$. Then we can write Assumptions MLR.1 and MLR.4 as

$$\text{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

and Assumption MLR.5 is the same as $\text{Var}(y|\mathbf{x}) = \sigma^2$. Stating the assumptions in this way clearly illustrates how Assumption MLR.5 differs greatly from Assumption MLR.4. Assumption MLR.4 says that the expected value of y, given $\mathbf{x}$, is linear in the parameters, but it certainly depends on $x_1, x_2, \ldots, x_k$. Assumption MLR.5 says that the variance of y, given $\mathbf{x}$, does *not* depend on the values of the independent variables.

We can now obtain the variances of the $\hat{\beta}_j$, where we again condition on the sample values of the independent variables. The proof is in the appendix to this chapter.

THEOREM 3.2 SAMPLING VARIANCES OF THE OLS SLOPE ESTIMATORS

Under Assumptions MLR.1 through MLR.5, conditional on the sample values of the independent variables,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j(1 - R_j^2)}, \qquad [3.51]$$

for $j = 1, 2, \ldots, k$, where $\text{SST}_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R-squared from regressing x_j on all other independent variables (and including an intercept).

The careful reader may be wondering whether there is a simple formula for the variance of $\hat{\beta}_j$ where we do not condition on the sample outcomes of the explanatory variables. The answer is: none that is useful. The formula in (3.51) is a highly nonlinear function of the x_{ij}, making averaging out across the population distribution of the explanatory variables virtually impossible. Fortunately, for any practical purpose equation (3.51) is what we want. Even when we turn to approximate, large-sample properties of OLS in Chapter 5 it turns out that (3.51) estimates the quantity we need for large-sample analysis, provided Assumptions MLR.1 through MLR.5 hold.

Before we study equation (3.51) in more detail, it is important to know that all of the Gauss-Markov assumptions are used in obtaining this formula. Whereas we did not need the homoskedasticity assumption to conclude that OLS is unbiased, we do need it to justify equation (3.51).

The size of $\text{Var}(\hat{\beta}_j)$ is practically important. A larger variance means a less precise estimator, and this translates into larger confidence intervals and less accurate hypotheses tests (as we will see in Chapter 4). In the next subsection, we discuss the elements comprising (3.51).

3-4a The Components of the OLS Variances: Multicollinearity

Equation (3.51) shows that the variance of $\hat{\beta}_j$ depends on three factors: σ^2, SST_j, and R_j^2. Remember that the index j simply denotes any one of the independent variables (such as education 3 or poverty rate). We now consider each of the factors affecting $\text{Var}(\hat{\beta}_j)$ in turn.

The Error Variance, σ^2. From equation (3.51), a larger σ^2 means larger sampling variances for the OLS estimators. This is not at all surprising: more "noise" in the equation (a larger σ^2) makes it more difficult to estimate the partial effect of any of the independent variables on y, and this is reflected in higher variances for the OLS slope estimators. Because σ^2 is a feature of the population, it has nothing to do with the sample size. It is the one component of (3.51) that is unknown. We will see later how to obtain an unbiased estimator of σ^2.

For a given dependent variable y, there is really only one way to reduce the error variance, and that is to add more explanatory variables to the equation (take some factors out of the error term). Unfortunately, it is not always possible to find additional legitimate factors that affect y.

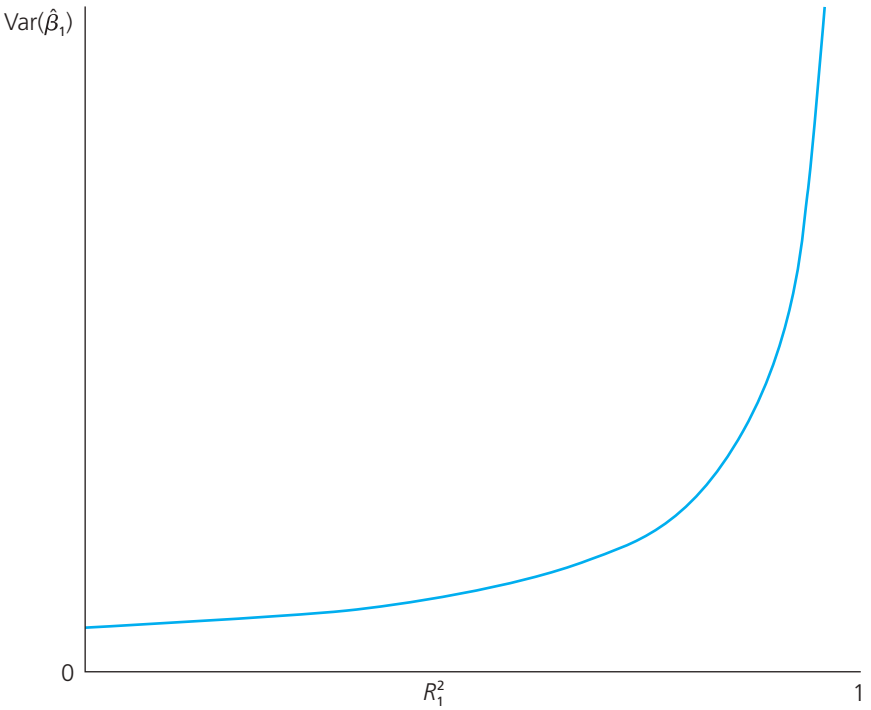
The Total Sample Variation in x_j, SST_j. From equation (3.51), we see that the larger the total variation in x_j is, the smaller is $\text{Var}(\hat{\beta}_j)$. Thus, everything else being equal, for estimating β_j we prefer to have as much sample variation in x_j as possible. We already discovered this in the simple regression case in Chapter 2. Although it is rarely possible for us to choose the sample values of the independent variables, there *is* a way to increase the sample variation in each of the independent variables: increase the sample size. In fact, when one randomly samples from a population, SST_j increases without bound as the sample size increases—roughly as a linear function of n. This is the component of the variance that systematically depends on the sample size.

When SST_j is small, $\text{Var}(\hat{\beta}_j)$ can get very large, but a small SST_j is not a violation of Assumption MLR.3. Technically, as SST_j goes to zero, $\text{Var}(\hat{\beta}_j)$ approaches infinity. The extreme case of no sample variation in x_j, $\text{SST}_j = 0$, is not allowed by Assumption MLR.3 because then we cannot even compute the OLS estimates.

The Linear Relationships among the Independent Variables, R_j^2. The term R_j^2 in equation (3.51) is the most difficult of the three components to understand. This term does not appear in simple regression analysis because there is only one independent variable in such cases. It is important to see that this R-squared is distinct from the R-squared in the regression of y on $x_1, x_2, \ldots, x_k$: R_j^2 is obtained from a regression involving only the independent variables in the original model, where x_j plays the role of a dependent variable.

Consider first the $k = 2$ case: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Then, $\text{Var}(\hat{\beta}_1) = \sigma^2/[\text{SST}_1(1 - R_1^2)]$, where R_1^2 is the R-squared from the simple regression of x_1 on x_2 (and an intercept, as always).

FIGURE 3.1 $\text{Var}(\hat{\beta}_1)$ as a function of R_1^2.



Because the *R*-squared measures goodness-of-fit, a value of R_1^2 close to one indicates that x_2 explains much of the variation in x_1 in the sample. This means that x_1 and x_2 are highly correlated.

As R_1^2 increases to one, $\text{Var}(\hat{\beta}_1)$ gets larger and larger. Thus, a high degree of linear relationship between x_1 and x_2 can lead to large variances for the OLS slope estimators. (A similar argument applies to $\hat{\beta}_2$.) See Figure 3.1 for the relationship between $\text{Var}(\hat{\beta}_1)$ and the *R*-squared from the regression of x_1 on x_2.

In the general case, R_j^2 is the proportion of the total variation in x_j that can be explained by the *other* independent variables appearing in the equation. For a given σ^2 and SST_j, the smallest $\text{Var}(\hat{\beta}_j)$ is obtained when $R_j^2 = 0$, which happens if, and only if, x_j has zero sample correlation with *every other* independent variable. This is the best case for estimating β_j, but it is rarely encountered.

The other extreme case, $R_j^2 = 1$, is ruled out by Assumption MLR.3, because $R_j^2 = 1$ means that, in the sample, x_j is a *perfect* linear combination of some of the other independent variables in the regression. A more relevant case is when R_j^2 is "close" to one. From equation (3.51) and Figure 3.1, we see that this can cause $\text{Var}(\hat{\beta}_j)$ to be large: $\text{Var}(\hat{\beta}_j) \to \infty$ as $R_j^2 \to 1$. High (but not perfect) correlation between two or more independent variables is called **multicollinearity**.

Before we discuss the multicollinearity issue further, it is important to be very clear on one thing: a case where R_j^2 is close to one is *not* a violation of Assumption MLR.3.

Because multicollinearity violates none of our assumptions, the "problem" of multicollinearity is not really well defined. When we say that multicollinearity arises for estimating β_j when R_j^2 is "close" to one, we put "close" in quotation marks because there is no absolute number that we can cite to conclude that multicollinearity is a problem. For example, $R_j^2 = .9$ means that 90% of the sample variation in x_j can be explained by the other independent variables in the regression model. Unquestionably, this means that x_j has a strong linear relationship to the other independent variables. But whether this translates into a $\text{Var}(\hat{\beta}_j)$ that is too large to be useful depends on the sizes of σ^2 and SST_j. As we will see in Chapter 4, for statistical inference, what ultimately matters is how big $\hat{\beta}_j$ is in relation to its standard deviation.

Just as a large value of R_j^2 can cause a large $\text{Var}(\hat{\beta}_j)$, so can a small value of SST_j. Therefore, a small sample size can lead to large sampling variances, too. Worrying about high degrees of correlation among the independent variables in the sample is really no different from worrying about a small sample size: both work to increase $\text{Var}(\hat{\beta}_j)$. The famous University of Wisconsin econometrician Arthur Goldberger, reacting to econometricians' obsession with multicollinearity, has (tongue in cheek) coined the term **micronumerosity**, which he defines as the "problem of small sample size." [For an engaging discussion of multicollinearity and micronumerosity, see Goldberger (1991).]

Although the problem of multicollinearity cannot be clearly defined, one thing is clear: everything else being equal, for estimating β_j, it is better to have less correlation between x_j and the other independent variables. This observation often leads to a discussion of how to "solve" the multicollinearity problem. In the social sciences, where we are usually passive collectors of data, there is no good way to reduce variances of unbiased estimators other than to collect more data. For a given data set, we can try dropping other independent variables from the model in an effort to reduce multicollinearity. Unfortunately, dropping a variable that belongs in the population model can lead to bias, as we saw in Section 3-3.

Perhaps an example at this point will help clarify some of the issues raised concerning multicollinearity. Suppose we are interested in estimating the effect of various school expenditure categories on student performance. It is likely that expenditures on teacher salaries, instructional materials, athletics, and so on are highly correlated: wealthier schools tend to spend more on everything, and poorer schools spend less on everything. Not surprisingly, it can be difficult to estimate the effect of any particular expenditure category on student performance when there is little variation in one category that cannot largely be explained by variations in the other expenditure categories (this leads to high R_j^2 for each of the expenditure variables). Such multicollinearity problems can be mitigated by collecting more data, but in a sense we have imposed the problem on ourselves: we are asking questions that may be too subtle for the available data to answer with any precision. We can probably do much better by changing the scope of the analysis and lumping all expenditure categories together, because we would no longer be trying to estimate the partial effect of each separate category.

Another important point is that a high degree of correlation between certain independent variables can be irrelevant as to how well we can estimate other parameters in the model. For example, consider a model with three independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

where x_2 and x_3 are highly correlated. Then $\text{Var}(\hat{\beta}_2)$ and $\text{Var}(\hat{\beta}_3)$ may be large. But the amount of correlation between x_2 and x_3 has no direct effect on $\text{Var}(\hat{\beta}_1)$. In fact, if x_1 is uncorrelated with x_2 and x_3, then $R_1^2 = 0$ and $\text{Var}(\hat{\beta}_1) = \sigma^2/\text{SST}_1$, regardless of how much correlation there is between x_2 and x_3. If β_1 is the parameter of interest, we do not really care about the amount of correlation between x_2 and x_3.

GOING FURTHER 3.4

Suppose you postulate a model explaining final exam score in terms of class attendance. Thus, the dependent variable is final exam score, and the key explanatory variable is number of classes attended. To control for student abilities and efforts outside the classroom, you include among the explanatory variables cumulative GPA, SAT score, and measures of high school performance. Someone says, "You cannot hope to learn anything from this exercise because cumulative GPA, SAT score, and high school performance are likely to be highly collinear." What should be your response?

The previous observation is important because economists often include many control variables in order to isolate the causal effect of a particular variable. For example, in looking at the relationship between loan approval rates and percentage of minorities in a neighborhood, we might include variables like average income, average housing value, measures of creditworthiness, and so on, because these factors need to be accounted for in order to draw causal conclusions about discrimination. Income, housing prices, and creditworthiness are generally highly correlated with each other.

But high correlations among these controls do not make it more difficult to determine the effects of discrimination.

Some researchers find it useful to compute statistics intended to determine the severity of multicollinearity in a given application. Unfortunately, it is easy to misuse such statistics because, as we have discussed, we cannot specify how much correlation among explanatory variables is "too much." Some multicollinearity "diagnostics" are omnibus statistics in the sense that they detect a strong linear relationship among any subset of explanatory variables. For reasons that we just saw, such statistics are of questionable value because they might reveal a "problem" simply because two control variables, whose coefficients we do not care about, are highly correlated. [Probably the most common omnibus multicollinearity statistic is the so-called *condition number*, which is defined in terms of the full data matrix and is beyond the scope of this text. See, for example, Belsley, Kuh, and Welsh (1980).]

Somewhat more useful, but still prone to misuse, are statistics for individual coefficients. The most common of these is the **variance inflation factor (VIF)**, which is obtained directly from equation (3.51). The VIF for slope coefficient j is simply $\text{VIF}_j = 1/(1 - R_j^2)$, precisely the term in $\text{Var}(\hat{\beta}_j)$ that is determined by correlation between x_j and the other explanatory variables. We can write $\text{Var}(\hat{\beta}_j)$ in equation (3.51) as

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j} \cdot \text{VIF}_j,$$

which shows that VIF_j is the factor by which $\text{Var}(\hat{\beta}_j)$ is higher because x_j is not uncorrelated with the other explanatory variables. Because VIF_j is a function of R_j^2—indeed, Figure 3.1 is essentially a graph of VIF_1—our previous discussion can be cast entirely in terms of the VIF. For example, if we had the choice, we would like VIF_j to be smaller (other things equal). But we rarely have the choice. If we think certain explanatory variables need to be included in a regression to infer causality of x_j, then we are hesitant to drop them, and whether we think VIF_j is "too high" cannot really affect that decision. If, say, our main interest is in the causal effect of x_1 on y, then we should ignore entirely the VIFs of other coefficients. Finally, setting a cutoff value for VIF above which we conclude multicollinearity is a "problem" is arbitrary and not especially helpful. Sometimes the value 10 is chosen: if VIF_j is above 10 (equivalently, R_j^2 is above .9), then we conclude that multicollinearity is a "problem" for estimating β_j. But a VIF_j above 10 does not mean that the standard deviation of $\hat{\beta}_j$ is too large to be useful because the standard deviation also depends on σ and SST_j, and the latter can be increased by increasing the sample size. Therefore, just as with looking at the size of R_j^2 directly, looking at the size of VIF_j is of limited use, although one might want to do so out of curiosity.

3-4b Variances in Misspecified Models

The choice of whether to include a particular variable in a regression model can be made by analyzing the tradeoff between bias and variance. In Section 3-3, we derived the bias induced by leaving out a relevant variable when the true model contains two explanatory variables. We continue the analysis of this model by comparing the variances of the OLS estimators.

Write the true population model, which satisfies the Gauss-Markov assumptions, as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

We consider two estimators of β_1. The estimator $\hat{\beta}_1$ comes from the multiple regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \qquad [3.52]$$

In other words, we include x_2, along with x_1, in the regression model. The estimator $\tilde{\beta}_1$ is obtained by omitting x_2 from the model and running a simple regression of y on x_1:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \qquad [3.53]$$

When $\beta_2 \neq 0$, equation (3.53) excludes a relevant variable from the model and, as we saw in Section 3-3, this induces a bias in $\tilde{\beta}_1$ unless x_1 and x_2 are uncorrelated. On the other hand, $\hat{\beta}_1$ is unbiased for β_1 for any value of β_2, including $\beta_2 = 0$. It follows that, if bias is used as the only criterion, $\hat{\beta}_1$ is preferred to $\tilde{\beta}_1$.

The conclusion that $\hat{\beta}_1$ is always preferred to $\tilde{\beta}_1$ does not carry over when we bring variance into the picture. Conditioning on the values of x_1 and x_2 in the sample, we have, from (3.51),

$$\text{Var}(\hat{\beta}_1) = \sigma^2/[\text{SST}_1(1 - R_1^2)], \tag{3.54}$$

where SST_1 is the total variation in x_1, and R_1^2 is the R-squared from the regression of x_1 on x_2. Further, a simple modification of the proof in Chapter 2 for two-variable regression shows that

$$\text{Var}(\tilde{\beta}_1) = \sigma^2/\text{SST}_1. \tag{3.55}$$

Comparing (3.55) to (3.54) shows that $\text{Var}(\tilde{\beta}_1)$ is always *smaller* than $\text{Var}(\hat{\beta}_1)$, unless x_1 and x_2 are uncorrelated in the sample, in which case the two estimators $\tilde{\beta}_1$ and $\hat{\beta}_1$ are the same. Assuming that x_1 and x_2 are not uncorrelated, we can draw the following conclusions:

1. When $\beta_2 \neq 0$, $\tilde{\beta}_1$ is biased, $\hat{\beta}_1$ is unbiased, and $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.
2. When $\beta_2 = 0$, $\tilde{\beta}_1$ and $\hat{\beta}_1$ are both unbiased, and $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.

From the second conclusion, it is clear that $\tilde{\beta}_1$ is preferred if $\beta_2 = 0$. Intuitively, if x_2 does not have a partial effect on y, then including it in the model can only exacerbate the multicollinearity problem, which leads to a less efficient estimator of β_1. A higher variance for the estimator of β_1 is the cost of including an irrelevant variable in a model.

The case where $\beta_2 \neq 0$ is more difficult. Leaving x_2 out of the model results in a biased estimator of β_1. Traditionally, econometricians have suggested comparing the likely size of the bias due to omitting x_2 with the reduction in the variance—summarized in the size of R_1^2—to decide whether x_2 should be included. However, when $\beta_2 \neq 0$, there are two favorable reasons for including x_2 in the model. The most important of these is that any bias in $\tilde{\beta}_1$ does not shrink as the sample size grows; in fact, the bias does not necessarily follow any pattern. Therefore, we can usefully think of the bias as being roughly the same for any sample size. On the other hand, $\text{Var}(\tilde{\beta}_1)$ and $\text{Var}(\hat{\beta}_1)$ both shrink to zero as n gets large, which means that the multicollinearity induced by adding x_2 becomes less important as the sample size grows. In large samples, we would prefer $\hat{\beta}_1$.

The other reason for favoring $\hat{\beta}_1$ is more subtle. The variance formula in (3.55) is conditional on the values of x_{i1} and x_{i2} in the sample, which provides the best scenario for $\tilde{\beta}_1$. When $\beta_2 \neq 0$, the variance of $\tilde{\beta}_1$ conditional only on x_1 is larger than that presented in (3.55). Intuitively, when $\beta_2 \neq 0$ and x_2 is excluded from the model, the error variance increases because the error effectively contains part of x_2. But the expression in equation (3.55) ignores the increase in the error variance because it will treat both regressors as nonrandom. For practical purposes, the σ^2 term in equation (3.55) increases when x_2 is dropped from the equation. A full discussion of the proper conditioning argument when computing the OLS variances would lead us too far astray. Suffice it to say that equation (3.55) is too generous when it comes to measuring the precision of $\tilde{\beta}_1$. Fortunately, statistical packages report the proper variance estimator, and so we need not worry about the subtleties in the theoretical formulas. After reading the next subsection, you might want to study Problems 14 and 15 for further insight.

3-4c Estimating σ^2: Standard Errors of the OLS Estimators

We now show how to choose an unbiased estimator of σ^2, which then allows us to obtain unbiased estimators of $\text{Var}(\hat{\beta}_j)$.

Because $\sigma^2 = \text{E}(u^2)$, an unbiased "estimator" of σ^2 is the sample average of the squared errors: $n^{-1}\sum_{i=1}^{n} u_i^2$. Unfortunately, this is not a true estimator because we do not observe the u_i. Nevertheless, recall that the errors can be written as $u_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik}$, and so the reason

we do not observe the u_i is that we do not know the β_j. When we replace each β_j with its OLS estimator, we get the OLS residuals:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}.$$

It seems natural to estimate σ^2 by replacing u_i with the $\hat{u}_i$. In the simple regression case, we saw that this leads to a biased estimator. The unbiased estimator of σ^2 in the general multiple regression case is

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2\right) \Big/ (n - k - 1) = \text{SSR}/(n - k - 1). \qquad [3.56]$$

We already encountered this estimator in the $k = 1$ case in simple regression.

The term $n - k - 1$ in (3.56) is the **degrees of freedom (*df*)** for the general OLS problem with n observations and k independent variables. Because there are $k + 1$ parameters in a regression model with k independent variables and an intercept, we can write

$$\begin{aligned} df &= n - (k + 1) \\ &= (\text{number of observations}) - (\text{number of estimated parameters}). \end{aligned} \qquad [3.57]$$

This is the easiest way to compute the degrees of freedom in a particular application: count the number of parameters, including the intercept, and subtract this amount from the number of observations. (In the rare case that an intercept is not estimated, the number of parameters decreases by one.)

Technically, the division by $n - k - 1$ in (3.56) comes from the fact that the expected value of the sum of squared residuals is $\text{E}(\text{SSR}) = (n - k - 1)\sigma^2$. Intuitively, we can figure out why the degrees of freedom adjustment is necessary by returning to the first order conditions for the OLS estimators. These can be written $\sum_{i=1}^{n} \hat{u}_i = 0$ and $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, where $j = 1, 2, \ldots, k$. Thus, in obtaining the OLS estimates, $k + 1$ restrictions are imposed on the OLS residuals. This means that, given $n - (k + 1)$ of the residuals, the remaining $k + 1$ residuals are known: there are only $n - (k + 1)$ degrees of freedom in the residuals. (This can be contrasted with the *errors* u_i, which have n degrees of freedom in the sample.)

For reference, we summarize this discussion with Theorem 3.3. We proved this theorem for the case of simple regression analysis in Chapter 2 (see Theorem 2.3). (A general proof that requires matrix algebra is provided in Advanced Treatment E.)

THEOREM 3.3 **UNBIASED ESTIMATION OF σ^2**

Under the Gauss-Markov assumptions MLR.1 through MLR.5, $\text{E}(\hat{\sigma}^2) = \sigma^2$.

The positive square root of $\hat{\sigma}^2$, denoted $\hat{\sigma}$, is called the **standard error of the regression (SER)**. The SER is an estimator of the standard deviation of the error term. This estimate is usually reported by regression packages, although it is called different things by different packages. (In addition to SER, $\hat{\sigma}$ is also called the *standard error of the estimate* and the *root mean squared error*.)

Note that $\hat{\sigma}$ can either decrease or increase when another independent variable is added to a regression (for a given sample). This is because, although SSR must fall when another explanatory variable is added, the degrees of freedom also falls by one. Because SSR is in the numerator and *df* is in the denominator, we cannot tell beforehand which effect will dominate.

For constructing confidence intervals and conducting tests in Chapter 4, we will need to estimate the **standard deviation of $\hat{\beta}_j$**, which is just the square root of the variance:

$$\text{sd}(\hat{\beta}_j) = \sigma/[\text{SST}_j(1 - R_j^2)]^{1/2}.$$

Because σ is unknown, we replace it with its estimator, $\hat{\sigma}$. This gives us the **standard error of $\hat{\beta}_j$**:

$$\text{se}(\hat{\beta}_j) = \hat{\sigma}/[\text{SST}_j(1 - R_j^2)]^{1/2}. \qquad [3.58]$$

Just as the OLS estimates can be obtained for any given sample, so can the standard errors. Because $\text{se}(\hat{\beta}_j)$ depends on $\hat{\sigma}$, the standard error has a sampling distribution, which will play a role in Chapter 4.

We should emphasize one thing about standard errors. Because (3.58) is obtained directly from the variance formula in (3.51), and because (3.51) relies on the homoskedasticity Assumption MLR.5, it follows that the standard error formula in (3.58) is *not* a valid estimator of $\text{sd}(\hat{\beta}_j)$ if the errors exhibit heteroskedasticity. Thus, while the presence of heteroskedasticity does not cause bias in the $\hat{\beta}_j$, it does lead to bias in the usual formula for $\text{Var}(\hat{\beta}_j)$, which then invalidates the standard errors. This is important because any regression package computes (3.58) as the default standard error for each coefficient (with a somewhat different representation for the intercept). If we suspect heteroskedasticity, then the "usual" OLS standard errors are invalid, and some corrective action should be taken. We will see in Chapter 8 what methods are available for dealing with heteroskedasticity.

For some purposes it is helpful to write

$$\text{se}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{n}\text{sd}(x_j)\sqrt{1 - R_j^2}}, \qquad [3.59]$$

in which we take $\text{sd}(x_j) = \sqrt{n^{-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$ to be the sample standard deviation where the total sum of squares is divided by n rather than $n - 1$. The importance of equation (3.59) is that it shows how the sample size, n, directly affects the standard errors. The other three terms in the formula—$\hat{\sigma}$, $\text{sd}(x_j)$, and R_j^2—will change with different samples, but as n gets large they settle down to constants. Therefore, we can see from equation (3.59) that the standard errors shrink to zero at the rate $1/\sqrt{n}$. This formula demonstrates the value of getting more data: the precision of the $\hat{\beta}_j$ increases as n increases. (By contrast, recall that unbiasedness holds for any sample size subject to being able to compute the estimators.) We will talk more about large sample properties of OLS in Chapter 5.

3-5 Efficiency of OLS: The Gauss-Markov Theorem

In this section, we state and discuss the important **Gauss-Markov Theorem**, which justifies the use of the OLS method rather than using a variety of competing estimators. We know one justification for OLS already: under Assumptions MLR.1 through MLR.4, OLS is unbiased. However, there are *many* unbiased estimators of the β_j under these assumptions (for example, see Problem 13). Might there be other unbiased estimators with variances smaller than the OLS estimators?

If we limit the class of competing estimators appropriately, then we can show that OLS *is* best within this class. Specifically, we will argue that, under Assumptions MLR.1 through MLR.5, the OLS estimator $\hat{\beta}_j$ for β_j is the **best linear unbiased estimator (BLUE)**. To state the theorem, we need to understand each component of the acronym "BLUE." First, we know what an estimator is: it is a rule that can be applied to any sample of data to produce an estimate. We also know what an unbiased estimator is: in the current context, an estimator, say, $\tilde{\beta}_1$, of β_j is an unbiased estimator of β_j if $\text{E}(\tilde{\beta}_j) = \beta_j$ for any $\beta_0, \beta_1, \ldots, \beta_k$.

What about the meaning of the term "linear"? In the current context, an estimator $\tilde{\beta}_j$ of β_j is linear if, and only if, it can be expressed as a linear function of the data on the dependent variable:

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i, \qquad [3.60]$$

where each w_{ij} can be a function of the sample values of all the independent variables. The OLS estimators are linear, as can be seen from equation (3.22).

Finally, how do we define "best"? For the current theorem, best is defined as *having the smallest variance*. Given two unbiased estimators, it is logical to prefer the one with the smallest variance (see Math Refresher C).

Now, let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, denote the OLS estimators in model (3.31) under Assumptions MLR.1 through MLR.5. The Gauss-Markov Theorem says that, for any estimator $\tilde{\beta}_j$ that is *linear* and *unbiased*, $\text{Var}(\hat{\beta}_j) \leq \text{Var}(\tilde{\beta}_j)$, and the inequality is usually strict. In other words, in the class of linear unbiased estimators, OLS has the smallest variance (under the five Gauss-Markov assumptions). Actually, the theorem says more than this. If we want to estimate any linear function of the β_j, then the corresponding linear combination of the OLS estimators achieves the smallest variance among all linear unbiased estimators. We conclude with a theorem, which is proven in Appendix 3A. It is because of this theorem that Assumptions MLR.1 through MLR.5 are known as the Gauss-Markov assumptions (for cross-sectional analysis).

THEOREM 3.4

GAUSS-MARKOV THEOREM

Under Assumptions MLR.1 through MLR.5, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the best linear unbiased estimators (BLUEs) of $\beta_0, \beta_1, \ldots, \beta_k$, respectively.

The importance of the Gauss-Markov Theorem is that, when the standard set of assumptions holds, we need not look for alternative unbiased estimators of the form in (3.60): none will be better than OLS. Equivalently, if we are presented with an estimator that is both linear and unbiased, then we know that the variance of this estimator is at least as large as the OLS variance; no additional calculation is needed to show this.

For our purposes, Theorem 3.4 justifies the use of OLS to estimate multiple regression models. If any of the Gauss-Markov assumptions fail, then this theorem no longer holds. We already know that failure of the zero conditional mean assumption (Assumption MLR.4) causes OLS to be biased, so Theorem 3.4 also fails. We also know that heteroskedasticity (failure of Assumption MLR.5) does not cause OLS to be biased. However, OLS no longer has the smallest variance among linear unbiased estimators in the presence of heteroskedasticity. In Chapter 8, we analyze an estimator that improves upon OLS when we know the brand of heteroskedasticity.

3-6 Some Comments on the Language of Multiple Regression Analysis

It is common for beginners, and not unheard of for experienced empirical researchers, to report that they "estimated an OLS model." Although we can usually figure out what someone means by this statement, it is important to understand that it is wrong—on more than just an aesthetic level—and reflects a misunderstanding about the components of a multiple regression analysis.

The first thing to remember is that ordinary least squares (OLS) is an estimation method, not a model. A model describes an underlying population and depends on unknown parameters. The *linear model* that we have been studying in this chapter can be written—in the population—as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \qquad [3.61]$$

where the parameters are the β_j. Importantly, we can talk about the meaning of the β_j without ever looking at data. It is true we cannot hope to learn much about the β_j without data, but the interpretation of the β_j is obtained from the linear model in equation (3.61).

Once we have a sample of data we can estimate the parameters. Although it is true that we have so far only discussed OLS as a possibility, there are actually many more ways to use the data than we can even list. We have focused on OLS due to its widespread use, which is justified by using the statistical considerations we covered previously in this chapter. But the various justifications for OLS rely on the assumptions we have made (MLR.1 through MLR.5). As we will see in later chapters, under

different assumptions different estimation methods are preferred—even though our model can still be represented by equation (3.61). Just a few examples include weighted least squares in Chapter 8, least absolute deviations in Chapter 9, and instrumental variables in Chapter 15.

One might argue that the discussion here is overly pedantic, and that the phrase "estimating an OLS model" should be taken as a useful shorthand for "I estimated a linear model by OLS." This stance has some merit, but we must remember that we have studied the properties of the OLS estimators under different assumptions. For example, we know OLS is unbiased under the first four Gauss-Markov assumptions, but it has no special efficiency properties without Assumption MLR.5. We have also seen, through the study of the omitted variables problem, that OLS is biased if we do not have Assumption MLR.4. The problem with using imprecise language is that it leads to vagueness on the most important considerations: what assumptions are being made on the underlying linear model? The issue of the assumptions we are using is conceptually different from the estimator we wind up applying.

Ideally, one writes down an equation like (3.61), with variable names that are easy to decipher, such as

$$math4 = \beta_0 + \beta_1 classize4 + \beta_2 math3 + \beta_3 \log(income) + \beta_4 motheduc + \beta_5 fatheduc + u \qquad [3.62]$$

if we are trying to explain outcomes on a fourth-grade math test. Then, in the context of equation (3.62), one includes a discussion of whether it is reasonable to maintain Assumption MLR.4, focusing on the factors that might still be in u and whether more complicated functional relationships are needed (a topic we study in detail in Chapter 6). Next, one describes the data source (which ideally is obtained via random sampling) as well as the OLS estimates obtained from the sample. A proper way to introduce a discussion of the estimates is to say "I estimated equation (3.62) by ordinary least squares. Under the assumption that no important variables have been omitted from the equation, and assuming random sampling, the OLS estimator of the class size effect, β_1, is unbiased. If the error term u has constant variance, the OLS estimator is actually best linear unbiased." As we will see in Chapters 4 and 5, we can often say even more about OLS. Of course, one might want to admit that while controlling for third-grade math score, family income and parents' education might account for important differences across students, it might not be enough—for example, u can include motivation of the student or parents—in which case OLS might be biased.

A more subtle reason for being careful in distinguishing between an underlying population model and an estimation method used to estimate a model is that estimation methods such as OLS can be used essentially as an exercise in curve fitting or prediction, without explicitly worrying about an underlying model and the usual statistical properties of unbiasedness and efficiency. For example, we might just want to use OLS to estimate a line that allows us to predict future college GPA for a set of high school students with given characteristics.

3-7 Several Scenarios for Applying Multiple Regression

Now that we have covered the algebraic and statistical properties of OLS, it is a good time to catalog different scenarios where unbiasedness of OLS can be established. In particular, we are interested in situations verifying Assumptions MLR.1 and MLR.4, as these are the important population assumptions. Assumption MLR.2 concerns the sampling scheme, and the mild restrictions in Assumption MLR.3 are rarely a concern.

The linearity assumption in MLR.1, where the error term u is additive, is always subject to criticism, although we know that it is not nearly as restrictive as it might seem because we can use transformations of both the explained and explanatory variables. Plus, the linear model is always a good starting point, and often provides a suitable approximation. In any case, for the purposes of the following discussion the functional form issue is not critical.

3-7a Prediction

As suggested at the end of Section 3-6, sometimes we are interested in a pure prediction exercise, where we hope to predict the outcome of a variable, y, given a set of observed variables, $x_1, x_2, \ldots, x_k$. To continue the example previously mentioned, a college admissions officer might want to predict the success of applicants—as measured by, say, future college GPA, y—based on information available at the time of application. These variables, which include performance variables from high school (GPA, kinds of classes taken, standardized test scores) and possibly family background, comprise the explanatory variables. As described in Math Refresher B, the best predictor of y, as measured by mean squared error, is the conditional expectation, $\mathrm{E}(y|x_1, \ldots, x_k)$. If we assume a linear function for the conditional expectation then

$$\mathrm{E}(y|x_1, \ldots, x_k) = \beta_0 = \beta_1 x_1 + \cdots + \beta_k x_k,$$

which is the same as writing

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$
$$\mathrm{E}(u|x_1, \ldots, x_k) = 0.$$

In other words, MLR.4 is true by construction once we assume linearity. If we have a random sample on the x_j and y and we can rule out perfect collinearity, we can obtain unbiased estimators of the β_j by OLS. In the example of predicting future GPAs, we would obtain a sample of students who attended the university so that we can observe their college GPAs. (Whether this provides a random sample from the relevant population is an interesting question, but too advanced to discuss here. We do so in Chapters 9 and 17.)

By estimating the β_j we can also see which factors are most important for predicting future college success—as a way to fine tune our prediction model. But we do not yet have a formal way of choosing which explanatory variables to include; that comes in the next chapter.

3-7b Efficient Markets

Efficient markets theories in economics, as well as some other theories, often imply that a single variable acts as a "sufficient statistic" for predicting the outcome variable, y. For emphasis, call this special predictor w. Then, given other observed factors, say $x_1, \ldots, x_k$, we might want to test the assumption

$$\mathrm{E}(y|w, \mathbf{x}) = \mathrm{E}(y|w), \qquad [3.63]$$

where $\mathbf{x}$ is a shorthand for $(x_1, x_2, \ldots, x_k)$. We can test (3.63) using a linear model for $\mathrm{E}(y|w, \mathbf{x})$:

$$\mathrm{E}(y|w, \mathbf{x}) = \beta_0 + \beta_1 w + \gamma_1 x_1 + \cdots + \gamma_k x_k, \qquad [3.64]$$

where the slight change in notation is used to reflect the special status of w. In Chapter 4 we will learn how to test whether all of the γ_j are zero:

$$\gamma_1 = \gamma_2 = \cdots = \gamma_k = 0. \qquad [3.65]$$

Many efficient markets theories imply more than just (3.63). In addition, typically

$$\mathrm{E}(y|w) = w,$$

which means that, in the linear equation (3.64), $\beta_0 = 0$ and $\beta_1 = 1$. Again, we will learn how to test such restrictions in Chapter 4.

As a specific example, consider the sports betting market—say, for college football. The gambling markets produce a point spread, $w = spread,$ which is determined prior to a game being played. The spread typically varies a bit during the days preceding the game, but it eventually settles on some

value. (Typically, the spread is in increments of 0.5.) The actually score differential in the game is $y = scorediff$. Efficiency of the gambling market implies that

$$\mathrm{E}(scorediff|spread, x_1, \ldots, x_k) = \mathrm{E}(scorediff|spread) = spread,$$

where $x_1, \ldots, x_k$ includes any variables observable to the public prior to the game being played. Examples include previous winning percentage, where the game is played, known injuries to key players, and so on. The idea is that, because lots of money is involved in betting, the spread will move until it incorporates all relevant information. Multiple regression can be used to test the efficient markets hypothesis because MLR.4 holds by construction once we assume a linear model:

$$y = \beta_0 + \beta_1 w + \gamma_1 x_1 + \cdots + \gamma_k x_k + u \quad [3.66]$$

$$\mathrm{E}(u|w, x_1, \ldots, x_k) = 0, \quad [3.67]$$

where the explanatory variables are $w, x_1, \ldots, x_k$.

Incidentally, it may be possible to think of a variable to be included among the x_j that the market has not incorporated into the spread. To be useful, it must be a variable that one can observe prior to the game being played. Most tests of the efficiency of the gambling markets show that, except for short aberrations, the market is remakably efficient.

3-7c Measuring the Tradeoff between Two Variables

Sometimes regression models are used not to predict, or to determe causality, but to simply measure how an economic agent trades off one variable for another. Call these variables y and w. For example, consider the population of K–12 teachers in a state in the United States. Let y be annual salary and w be a measure of pension compensation. If teachers are indifferent between a dollar of salary and a dollar of pension, then, on average, a one-dollar increase in pension compensation should be associated with a one-dollar fall in salary. In other words, only total compensation matters. Naturally, this is a ceteris paribus question: all other relevant factors should be held fixed. In particular, we would expect to see a positive correlation between salary and pension benefits because pension benefits are often tied to salary. We want to know, for a given teacher, how does that teacher trade off one for the other.

Because we are simply measuring a tradeoff, it should not matter which variable we choose as y and which we choose as w. However, functional form considerations can come into play. (We will see this later in Example 4.10 in Chapter 4, where we study the salary-benefits tradeoff using aggregated data.) Once we have chosen y and w, and we have controls $\mathbf{x} = (x_1, \ldots, x_k)$, we are, as in Section 3-7b, interested in $\mathrm{E}(y|w, \mathbf{x})$. Assuming a linear model, we are exactly in the situation given in equations (3.66) and (3.67). A key difference is that, assuming the x_j properly control for differences in individuals, the theory of a one-to-one tradeoff is $\beta_1 = -1$, without restricting the intercept, β_0. That is quite different from the efficient markets hypothesis. Further, we include the x_j to control for differences; we do not expect the y_j to be zero, and we would generally have no interest in testing (3.65).

If we are not able to include sufficient controls in $\mathbf{x}$ then the estimated tradeoff coefficient, β, will be biased (although the direction depends on what we think we have omitted). This is tantamount to an omitted variable problem. For example, we may not have a suitable measure of teachers' taste for saving or amount of risk aversion.

3-7d Testing for Ceteris Paribus Group Differences

Another common application of multiple regression analysis is to test for differences among groups—often, groups of people—once we account for other factors. In Section 2-7 we discussed the example of estimating differences in hourly wage, *wage,* based on race, which is divided into white and other. To this end, define a binary variable *white*. In Section 2-7a we noted that finding a difference in average wages across whites and nonwhites did not necessarily indicate wage discrimination because other factors could contribute to such a differece.

Let $x_1, x_2, \ldots, x_k$ denote other observable factors that can affect hourly wage—such as education, workforce experience, and so on. Then we are interested in

$$\mathrm{E}(wage|white, x_1, \ldots, x_k).$$

If we have accounted for all factors in wage that should affect productivity, then *wage* differences by race might be attributable to discrimination. In the simplest case, we would use a linear model:

$$\mathrm{E}(wage|white, x_1, \ldots, x_k) = \beta_0 + \beta_1 white + \gamma_1 x_1 + \cdots + \gamma_k x_k, \quad [3.68]$$

where we are primarily interested in the coefficient β, which measures the difference in whites and nonwhites given the *same* levels of the control variables, $x_1, x_2, \ldots, x_k$ (education, experience, and so on). For a general y and w, we again have (3.66) and (3.67) in force, and so MLR.4 holds by construction. OLS can be used to obtain an unbiased estimator of β (and the other coefficients). Problems arise when we cannot include all suitable variables among the x_j, in which case, again, we have an omitted variable problem. In the case of testing for racial or gender discrimination, failure to control for all relevant factors can cause systematic bias in estimating discrepancies due to discrimination.

3-7e Potential Outcomes, Treatment Effects, and Policy Analysis

For most practicing economists, the most exciting applications of multiple regression are in trying to estimate causal effects of policy interventions. Do job training programs increase labor earnings? By how much? Do school choice programs improve student outcomes? Does legalizing marijuana increase crime rates?

We introduced the potential outcomes approach to studying policy questions in Section 2-7a. In particular, we studied simple regression in the context of a binary policy intervention, using the notion of counterfactual outcomes. In this section we change notation slightly, using w to denote the binary intervention or policy indicator. As in Section 2-7a, for each unit in the population we imagine the existence of the potential outcomes, $y(0)$ and $y(1)$—representing different states of the world. If we assume a constant treatment effect, say τ, then we can write, for any unit i,

$$y_i(1) = \tau + y_i(0).$$

When the treatment effect can vary by i, the average treatment effect is

$$\tau_{ate} = \mathrm{E}[y_i(1) - y_i(0)], \quad [3.69]$$

where the expectation is taken over the entire population.

For a random draw i, the outcome we observe, y_i, can be written

$$y_i = (1 - w_i)y_i(0) + w_i y_i(1). \quad [3.70]$$

One of the important conclusions from Section 2-7a is that the simple regression of y on w (with an intercept, as usual) is an unbiased estimator of τ_{ate} only if we have random assignment of w—that is,

$$w \text{ is independent of } [y(0), y(1)].$$

Random assignment is still pretty rare in business, economics, and other social sciences because true experiments are still somewhat rare. Fortunately, if we can control variables—variables that help predict the potential outcomes and determine assignment into the treatment and control groups—we can use multiple regression. Letting $\mathbf{x}$ again denote a set of control variables, consider the following assumption:

$$w \text{ is independent of } [y(0), y(1)] \text{ conditional on } \mathbf{x}. \quad [3.71]$$

For fairly obvious reasons, this assumption is called **conditional independence**, where it is important to note the variables in $\mathbf{x}$ that are in the conditioning set. In the treatment effects literature,

(3.71) is also called **unconfounded assignment** or *unconfoundedness conditional on* **x**. The terms **ignorable assignment** and *ignorability* are also used.

Assumption (3.71) has a simple interpretation. Think of partitioning the population based on the observed variables in **x**. For concreteness, consider the job training program introduced in Section 2-7a. There, w indicates whether a worker participates in a job training program, and y is an outcome such as labor income. The elements in **x** include education, age, and past labor market history, such as earnings from the previous couple of years. Suppose that workers are more likely to participate in the program the lower their education, age, and the worse their previous labor market outcomes. Then, because education, age, and prior labor market history are very likely to predict $y(0)$ and $y(1)$, random assignment does not hold. Nevertheless, once we group people by education, age, and prior work history, it is possible that assignment is random. As a concrete example, consider the group of people with 12 years of schooling who are 35 years old and who had average earnings of \$25,000 the past two years. What (3.71) requires is that within this group, assignment to the treatment and control groups is random.

The more variables we observe prior to implementation of the program the more likely (3.71) is to hold. If we observe no information to include in **x** then we are back to assuming pure random assignment. Of course, it is always possible that we have not included the correct variables in **x**. For example, perhaps everyone in a sample from the eligible population was administered a test to measure intelligence, and assignment to the program is partly based on the score from the test. If we observe the test score, we include it in **x**. If we cannot observe the test score, it must be excluded from **x** and (3.71) would generally fail—although it could be "close" to being true if we have other good controls in **x**.

How can we use (3.71) in multiple regression? Here we only consider the case of a constant treatment effect, τ. Section 7-6 in Chapter 7 considers the more general case. Then, in the population,

$$y = y(0) + \tau w$$

and

$$\mathrm{E}(y|w, \mathbf{x}) = \mathrm{E}[y(0)|w, \mathbf{x}] + \tau w = \mathrm{E}[y(0)|\mathbf{x}] + \tau w, \qquad [3.72]$$

where the second equality follows from conditional independence. Now assume that $\mathrm{E}[y(0)|\mathbf{x}]$ is linear,

$$\mathrm{E}[y(0)|\mathbf{x}] = \alpha + \mathbf{x}\boldsymbol{\gamma}.$$

Plugging in gives

$$\mathrm{E}(y|w, \mathbf{x}) = \alpha + \tau w + \mathbf{x}\boldsymbol{\gamma} = \alpha + \tau w + \gamma_1 x_1 + \cdots + x_k x_k. \qquad [3.73]$$

As in several previous examples in this section, we are interested primarily in the coefficient on w, which we have called τ. The γ_j are of interest for logical consistency checks—for example, we should expect more education to lead to higher earnings, on average—but the main role of the x_j is to control for differences across units.

In Chapter 7 we will cover treatment effects in more generality, including how to use multiple regression when treatment effects vary by unit (individual in the job training case).

EXAMPLE 3.7 Evaluating a Job Training Program

The data in JTRAIN98 are on male workers that can be used to evaluate a job training program, where the variable we would like to explain, $y = earn98$ is labor market earnings in 1998, the year following the job training program (which took place in 1997). The earnings variable is measured in thousands of dollars. The variable $w = train$ is the binary participation (or "treatment") indicator. The participation in the job training program was partly based on past labor market outcomes and

is partly voluntary. Therefore, random assignment is unlikely to be a good assumption. As control variables we use earnings in 1996 (the year prior to the program), years of schooling (*educ*), age, and marital status (*married*). Like the training indicator, marital status is coded as a binary variable, where *married* = 1 means the man is married.

The simple regression estimates are

$$\widehat{earn98} = 10.61 - 2.05\ train \qquad [3.74]$$

$$n = 1{,}130,\ R^2 = 0.016$$

Because *earns*98 is measured in thousands of dollars, the coefficient on *train,* −2.05, shows that, on average, those participating in the program earned $2,050 *less* than those who did not. The average earnings for those who did not participate is gotten from the intercept, so $10,610.

Without random assignment, it is possible, even likely, that the negative (and large in magnitude) coefficient on *train* is a product of nonrandom selection into participation. This could be either because men with poor earnings histories were more likely to be chosen or that such men are more likely to participate if made eligible. We will not examine these propositions in detail here. Instead, we add the four controls and perform a multiple regression:

$$\widehat{earn98} = 4.67 + 2.41\ train + .373\ earn96 + .363\ educ - .181\ age + 2.48\ married \qquad [3.75]$$

$$n = 1{,}130,\ R^2 = 0.405$$

The change in the coefficient on *train* is remarkable: the program is now estimated to increase earnings, on average, by $2,410. In other words, controlling for differences in preprogram earnings, education levels, age, and marital status produces a much different estimate than the simple difference-in-means estimate.

The signs of the coefficients on the control variables are not surprising. We expect earnings to be positively correlated over time—so *earns*96 has a positive coefficient. Workers with more education also earn more: about $363 for each additional year. The marriage effect is roughly as large as the job training effect: ceteris paribus, married men earn, on average, about $2,480 more than their single counterparts.

The predictability of the control variables is indicated by the *R*-squared in the multiple regression, $R^2 = 0.405$. There is still much unexplained variation, but collectively the variables do a pretty good job.

GOING FURTHER 3.5

Does it make sense to compare the intercepts in equations (3.74) and (3.75)? Explain.

Before we end this section, a final remark is in order. As with the other examples in this chapter, we have not determined the statistical significance of the estimates. We remedy this omission in Chapter 4, where we learn how to test whether there is an effect in the entire population, and also obtain confidence intervals for the parameters, such as the average treatment effect of a job training program.

Summary

1. The multiple regression model allows us to effectively hold other factors fixed while examining the effects of a particular independent variable on the dependent variable. It explicitly allows the independent variables to be correlated.
2. Although the model is linear in its *parameters*, it can be used to model nonlinear relationships by appropriately choosing the dependent and independent variables.

3. The method of ordinary least squares is easily applied to estimate the multiple regression model. Each slope estimate measures the partial effect of the corresponding independent variable on the dependent variable, holding all other independent variables fixed.
4. R^2 is the proportion of the sample variation in the dependent variable explained by the independent variables, and it serves as a goodness-of-fit measure. It is important not to put too much weight on the value of R^2 when evaluating econometric models.
5. Under the first four Gauss-Markov assumptions (MLR.1 through MLR.4), the OLS estimators are unbiased. This implies that including an irrelevant variable in a model has no effect on the unbiasedness of the intercept and other slope estimators. On the other hand, omitting a relevant variable causes OLS to be biased. In many circumstances, the direction of the bias can be determined.
6. Under the five Gauss-Markov assumptions, the variance of an OLS slope estimator is given by $\text{Var}(\hat{\beta}_j) = \sigma^2/[\text{SST}_j(1 - R_j^2)]$. As the error variance σ^2 increases, so does $\text{Var}(\hat{\beta}_j)$, while $\text{Var}(\hat{\beta}_j)$ decreases as the sample variation in x_j, SST_j, increases. The term R_j^2 measures the amount of collinearity between x_j and the other explanatory variables. As R_j^2 approaches one, $\text{Var}(\hat{\beta}_j)$ is unbounded.
7. Adding an irrelevant variable to an equation generally increases the variances of the remaining OLS estimators because of multicollinearity.
8. Under the Gauss-Markov assumptions (MLR.1 through MLR.5), the OLS estimators are the best linear unbiased estimators (BLUEs).
9. Section 3-7 discusses the various ways that multiple regression analysis is used in economics and other social sciences, including for prediction, testing efficient markets, estimating tradeoffs between variables, and evaluating policy interventions. We will see examples of all such applications in the remainder of the text.
10. Beginning in Chapter 4, we will use the standard errors of the OLS coefficients to compute confidence intervals for the population parameters and to obtain test statistics for testing hypotheses about the population parameters. Therefore, in reporting regression results we now include the standard errors along with the associated OLS estimates. In equation form, standard errors are usually put in parentheses below the OLS estimates, and the same convention is often used in tables of OLS output.

THE GAUSS-MARKOV ASSUMPTIONS

The following is a summary of the five Gauss-Markov assumptions that we used in this chapter. Remember, the first four were used to establish unbiasedness of OLS, whereas the fifth was added to derive the usual variance formulas and to conclude that OLS is best linear unbiased.

Assumption MLR.1 (Linear in Parameters)

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term.

Assumption MLR.2 (Random Sampling)

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

Assumption MLR.3 (No Perfect Collinearity)

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

Assumption MLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any values of the independent variables. In other words,

$$\text{E}(u|x_1, x_2, \ldots, x_k) = 0.$$

Assumption MLR.5 (Homoskedasticity)

The error u has the same variance given any value of the explanatory variables. In other words,

$$\text{Var}(u|x_1, \ldots, x_k) = \sigma^2.$$

Key Terms

Best Linear Unbiased Estimator (BLUE)
Biased Toward Zero
Ceteris Paribus
Conditional Independence
Degrees of Freedom (*df*)
Disturbance
Downward Bias
Endogenous Explanatory Variable
Error Term
Excluding a Relevant Variable
Exogenous Explanatory Variables
Explained Sum of Squares (SSE)
First Order Conditions
Frisch-Waugh Theorem
Gauss-Markov Assumptions
Gauss-Markov Theorem
Ignorable Assignment
Inclusion of an Irrelevant Variable
Intercept
Micronumerosity
Misspecification Analysis
Multicollinearity
Multiple Linear Regression (MLR) Model
Multiple Regression Analysis
OLS Intercept Estimate
OLS Regression Line
OLS Slope Estimates
Omitted Variable Bias
Ordinary Least Squares
Overspecifying the Model
Partial Effect
Perfect Collinearity
Population Model
Residual
Residual Sum of Squares
Sample Regression Function (SRF)
Slope Parameters
Standard Deviation of $\hat{\beta}_j$
Standard Error of $\hat{\beta}_j$
Standard Error of the Regression (SER)
Sum of Squared Residuals (SSR)
Total Sum of Squares (SST)
True Model
Unconfounded Assignment
Underspecifying the Model
Upward Bias
Variance Inflation Factor (VIF)

Problems

1 Using the data in GPA2 on 4,137 college students, the following equation was estimated by OLS:

$$\widehat{colgpa} = 1.392 - .0135\, hsperc + .00148\, sat$$
$$n = 4{,}137,\ R^2 = .273,$$

where *colgpa* is measured on a four-point scale, *hsperc* is the percentile in the high school graduating class (defined so that, for example, *hsperc* = 5 means the *top* 5% of the class), and *sat* is the combined math and verbal scores on the student achievement test.

(i) Why does it make sense for the coefficient on *hsperc* to be negative?

(ii) What is the predicted college GPA when *hsperc* = 20 and *sat* = 1,050?

(iii) Suppose that two high school graduates, A and B, graduated in the same percentile from high school, but Student A's SAT score was 140 points higher (about one standard deviation in the sample). What is the predicted difference in college GPA for these two students? Is the difference large?

(iv) Holding *hsperc* fixed, what difference in SAT scores leads to a predicted *colgpa* difference of .50, or one-half of a grade point? Comment on your answer.

2 The data in WAGE2 on working men was used to estimate the following equation:

$$\widehat{educ} = 10.36 - .094\, sibs + .131\, meduc + .210\, feduc$$
$$n = 722,\ R^2 = .214,$$

where *educ* is years of schooling, *sibs* is number of siblings, *meduc* is mother's years of schooling, and *feduc* is father's years of schooling.

(i) Does *sibs* have the expected effect? Explain. Holding *meduc* and *feduc* fixed, by how much does *sibs* have to increase to reduce predicted years of education by one year? (A noninteger answer is acceptable here.)

(ii) Discuss the interpretation of the coefficient on *meduc*.

(iii) Suppose that Man A has no siblings, and his mother and father each have 12 years of education, and Man B has no siblings, and his mother and father each have 16 years of education. What is the predicted difference in years of education between B and A?

3 The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* and *age* are measured in years. (See also Computer Exercise C3 in Chapter 2.)

(i) If adults trade off sleep for work, what is the sign of β_1?

(ii) What signs do you think β_2 and β_3 will have?

(iii) Using the data in SLEEP75, the estimated equation is

$$\widehat{sleep} = 3{,}638.25 - .148\, totwrk - 11.13\, educ + 2.20\, age$$
$$n = 706,\ R^2 = .113.$$

If someone works five more hours per week, by how many minutes is *sleep* predicted to fall? Is this a large tradeoff?

(iv) Discuss the sign and magnitude of the estimated coefficient on *educ*.

(v) Would you say *totwrk*, *educ*, and *age* explain much of the variation in *sleep*? What other factors might affect the time spent sleeping? Are these likely to be correlated with *totwrk*?

4 The median starting salary for new law school graduates is determined by

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u,$$

where *LSAT* is the median LSAT score for the graduating class, *GPA* is the median college GPA for the class, *libvol* is the number of volumes in the law school library, *cost* is the annual cost of attending law school, and *rank* is a law school ranking (with *rank* = 1 being the best).

(i) Explain why we expect $\beta_5 \leq 0$.

(ii) What signs do you expect for the other slope parameters? Justify your answers.

(iii) Using the data in LAWSCH85, the estimated equation is

$$\widehat{\log(salary)} = 8.34 + .0047\, LSAT + .248\, GPA + .095 \log(libvol) + .038 \log(cost) - .0033\, rank$$
$$n = 136,\ R^2 = .842.$$

What is the predicted ceteris paribus difference in salary for schools with a median GPA different by one point? (Report your answer as a percentage.)

(iv) Interpret the coefficient on the variable log(*libvol*).

(v) Would you say it is better to attend a higher ranked law school? How much is a difference in ranking of 20 worth in terms of predicted starting salary?

5 In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

(i) In the model

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u,$$

does it make sense to hold *sleep*, *work*, and *leisure* fixed, while changing *study*?

(ii) Explain why this model violates Assumption MLR.3.

(iii) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.3?

6 Consider the multiple regression model containing three independent variables, under Assumptions MLR.1 through MLR.4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

You are interested in estimating the sum of the parameters on x_1 and x_2; call this $\theta_1 = \beta_0 + \beta_1$.

(i) Show that $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ is an unbiased estimator of θ_1.

(ii) Find $\text{Var}(\hat{\theta}_1)$ in terms of $\text{Var}(\hat{\beta}_1)$,$\text{Var}(\hat{\beta}_2)$, and $\text{Corr}(\hat{\beta}_1, \hat{\beta}_2)$.

7 Which of the following can cause OLS estimators to be biased?

(i) Heteroskedasticity.

(ii) Omitting an important variable.

(iii) A sample correlation coefficient of .95 between two independent variables both included in the model.

8 Suppose that average worker productivity at manufacturing firms (*avgprod*) depends on two factors, average hours of training (*avgtrain*) and average worker ability (*avgabil*):

$$avgprod = \beta_0 + \beta_1 avgtrain + \beta_2 avgabil + u.$$

Assume that this equation satisfies the Gauss-Markov assumptions. If grants have been given to firms whose workers have less than average ability, so that *avgtrain* and *avgabil* are negatively correlated, what is the likely bias in $\tilde{\beta}_1$ obtained from the simple regression of *avgprod* on *avgtrain*?

9 The following equation describes the median housing price in a community in terms of amount of pollution (*nox* for nitrous oxide) and the average number of rooms in houses in the community (*rooms*):

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u.$$

(i) What are the probable signs of β_1 and β_2? What is the interpretation of β_1? Explain.

(ii) Why might *nox* [or more precisely, log(*nox*)] and *rooms* be negatively correlated? If this is the case, does the simple regression of log(*price*) on log(*nox*) produce an upward or a downward biased estimator of β_1?

(iii) Using the data in HPRICE2, the following equations were estimated:

$$\widehat{\log(price)} = 11.71 - 1.043 \log(nox), n = 506, R^2 = .264.$$

$$\widehat{\log(price)} = 9.23 - .718 \log(nox) + .306 rooms, n = 506, R^2 = .514.$$

Is the relationship between the simple and multiple regression estimates of the elasticity of *price* with respect to *nox* what you would have predicted, given your answer in part (ii)? Does this mean that $-.718$ is definitely closer to the true elasticity than -1.043?

10 Suppose that you are interested in estimating the ceteris paribus relationship between y and x_1. For this purpose, you can collect data on two control variables, x_2 and x_3. (For concreteness, you might think of y as final exam score, x_1 as class attendance, x_2 as GPA up through the previous semester, and x_3 as SAT or ACT score.) Let $\tilde{\beta}_1$ be the simple regression estimate from y on x_1 and let $\hat{\beta}_1$ be the multiple regression estimate from y on x_1, x_2, x_3.

(i) If x_1 is highly correlated with x_2 and x_3 in the sample, and x_2 and x_3 have large partial effects on y, would you expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be similar or very different? Explain.

(ii) If x_1 is almost uncorrelated with x_2 and x_3, but x_2 and x_3 are highly correlated, will $\tilde{\beta}_1$ and $\hat{\beta}_1$ tend to be similar or very different? Explain.

(iii) If x_1 is highly correlated with x_2 and x_3, and x_2 and x_3 have small partial effects on y, would you expect $\text{se}(\tilde{\beta}_1)$ or $\text{se}(\hat{\beta}_1)$ to be smaller? Explain.

(iv) If x_1 is almost uncorrelated with x_2 and x_3, x_2 and x_3 have large partial effects on y, and x_2 and x_3 are highly correlated, would you expect $\text{se}(\tilde{\beta}_1)$ or $\text{se}(\hat{\beta}_1)$ to be smaller? Explain.

11 Suppose that the population model determining y is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

and this model satisfies Assumptions MLR.1 through MLR.4. However, we estimate the model that omits x_3. Let $\tilde{\beta}_0$, $\tilde{\beta}_1$, and $\tilde{\beta}_2$ be the OLS estimators from the regression of y on x_1 and x_2. Show that the expected value of $\tilde{\beta}_1$ (given the values of the independent variables in the sample) is

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2},$$

where the $\hat{r}_{i1}$ are the OLS residuals from the regression of x_1 on x_2. [*Hint:* The formula for $\tilde{\beta}_1$ comes from equation (3.22). Plug $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$ into this equation. After some algebra, take the expectation treating x_{i3} and $\hat{r}_{i1}$ as nonrandom.]

12 The following equation represents the effects of tax revenue mix on subsequent employment growth for the population of counties in the United States:

$$growth = \beta_0 + \beta_1 share_{\mathrm{P}} + \beta_2 share_{\mathrm{I}} + \beta_3 share_{\mathrm{S}} + other\ factors,$$

where *growth* is the percentage change in employment from 1980 to 1990, $share_{\mathrm{P}}$ is the share of property taxes in total tax revenue, $share_{\mathrm{I}}$ is the share of income tax revenues, and $share_{\mathrm{S}}$ is the share of sales tax revenues. All of these variables are measured in 1980. The omitted share, $share_{\mathrm{F}}$, includes fees and miscellaneous taxes. By definition, the four shares add up to one. Other factors would include expenditures on education, infrastructure, and so on (all measured in 1980).

(i) Why must we omit one of the tax share variables from the equation?

(ii) Give a careful interpretation of β_1.

13 (i) Consider the simple regression model $y = \beta_0 + \beta_1 x + u$ under the first four Gauss-Markov assumptions. For some function $g(x)$, for example $g(x) = x^2$ or $g(x) = \log(1 + x^2)$, define $z_i = g(x_i)$. Define a slope estimator as

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n} (z_i - \bar{z}) y_i\right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z}) x_i\right).$$

Show that $\tilde{\beta}_1$ is linear and unbiased. Remember, because $\mathrm{E}(u|x) = 0$, you can treat both x_i and z_i as nonrandom in your derivation.

(ii) Add the homoskedasticity assumption, MLR.5. Show that

$$\mathrm{Var}(\tilde{\beta}_1) = \sigma^2 \left(\sum_{i=1}^{n} (z_i - \bar{z})^2\right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z}) x_i\right)^2.$$

(iii) Show directly that, under the Gauss-Markov assumptions, $\mathrm{Var}(\hat{\beta}_1) \leq \mathrm{Var}(\tilde{\beta}_1)$, where $\hat{\beta}_1$ is the OLS estimator. [*Hint:* The Cauchy-Schwartz inequality in Math Refresher B implies that

$$\left(n^{-1} \sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})\right)^2 \leq \left(n^{-1} \sum_{i=1}^{n} (z_i - \bar{z})^2\right)\left(n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2\right);$$

notice that we can drop $\bar{x}$ from the sample covariance.]

14 Suppose you have a sample of size n on three variables, y, x_1, and x_2, and you are primarily interested in the effect of x_1 on y. Let $\tilde{\beta}_1$ be the coefficient on x_1 from the simple regression and $\hat{\beta}_1$ the coefficient on x_1 from the regression y on x_1, x_2. The standard errors reported by any regression package are

$$\mathrm{se}(\tilde{\beta}_1) = \frac{\tilde{\sigma}}{\sqrt{\mathrm{SST}_1}}$$

$$\text{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\text{SST}_1}} \cdot \sqrt{\text{VIF}_1},$$

where $\tilde{\sigma}$ is the SER from the simple regression, $\hat{\sigma}$ is the SER from the multiple regression, $\text{VIF}_1 = 1/(1 - R_1^2)$, and R_1^2 is the R-squared from the regression of x_1 on x_2. Explain why $\text{se}(\hat{\beta}_1)$ can be smaller or larger than $\text{se}(\tilde{\beta}_1)$.

15 The following estimated equations use the data in MLB1, which contains information on major league baseball salaries. The dependent variable, *lsalary*, is the log of salary. The two explanatory variables are years in the major leagues (*years*) and runs batted in per year (*rbisyr*):

$$\widehat{lsalary} = 12.373 + .1770\ years$$
$$(.098)\ (.0132)$$
$$n = 353,\ \text{SSR} = 326.196,\ \text{SER} = .964,\ R^2 = .337$$

$$\widehat{lsalary} = 11.861 + .0904\ years + .0302\ rbisyr$$
$$(.084)\ (.0118)\ \ (.0020)$$
$$n = 353,\ \ \text{SSR} = 198.475,\ \text{SER} = .753,\ R^2 = .597$$

(i) How many degrees of freedom are in each regression? Why is the SER smaller in the second regression than the first?

(ii) The sample correlation coefficient between *years* and *rbisyr* is about 0.487. Does this make sense? What is the variance inflation factor (there is only one) for the slope coefficients in the multiple regression? Would you say there is little, moderate, or strong collinearity between *years* and *rbisyr*?

(iii) How come the standard error for the coefficient on *years* in the multiple regression is lower than its counterpart in the simple regression?

16 The following equations were estimated using the data in LAWSCH85:

$$\widehat{lsalary} = 9.90 - .0041\ rank + .294\ GPA$$
$$(.24)\ (.0003)\ \ (.069)$$
$$n = 142,\ R^2 = .8238$$

$$\widehat{lsalary} = 9.86 - .0038\ rank + .295\ GPA + .00017\ age$$
$$(.29)\ (.0004)\ \ (.083)\ \ (.00036)$$
$$n = 99,\ R^2 = .8036$$

How can it be that the R-squared is smaller when the variable *age* is added to the equation?

17 Consider an estimated equation for workers earning an hourly wage, *wage*,where *educ,* years of schooling, and *exper,* actual years in the workforce, are measured in years. The dependent variable is *lwage* = log(*wage*)*:*

$$\widehat{lwage} = 0.532 + .094\ educ + .026\ exper$$
$$n = 932,\ R^2 = 0.188$$

Suppose that getting one more year of education necessarily reduces workforce experience by one year. What is the estimated percentage change in *wage* from getting one more year of schooling?

18 The potential outcomes framework in Section 3-7e can be extended to more than two potential outcomes. In fact, we can think of the policy variable, w, as taking on many different values, and then $y(w)$ denotes the outcome for policy level w. For concreteness, suppose w is the dollar amount of a grant that can be used for purchasing books and electronics in college, $y(w)$ is a measure of college performance,

such as grade point average. For example, $y(0)$ is the resulting GPA if the student receives no grant and $y(500)$ is the resulting GPA if the grant amount is \$500.

For a random draw i, we observe the grant level, $w_i \geq 0$ and $y_i = y(w_i)$. As in the binary program evaluation case, we observe the policy level, w_i, and then only the outcome associated with that level.

(i) Suppose a linear relationship is assumed:

$$y(w) = \alpha + \beta w + v(0)$$

where $y(0) = \alpha + v$. Further, assume that for all i, w_i is independent of v_i. Show that for each i we can write

$$y_i = \alpha + \beta w_i + v_i$$
$$\mathrm{E}(v_i|w_i) = 0.$$

(ii) In the setting of part (i), how would you estimate β (and α) given a random sample? Justify your answer.

(iii) Now suppose that w_i is possibly correlated with v_i, but for a set of observed variables x_{ij},

$$\mathrm{E}(v_i|w_i, x_{i1}, \ldots, x_{ik}) = \mathrm{E}(v_i|x_{i1}, \ldots, x_{ik}) = \eta + \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik}.$$

The first equality holds if w_i is independent of v_i conditional on $(x_{i1}, \ldots, x_{ik})$ and the second equality assumes a linear relationship. Show that we can write

$$y_i = \psi + \beta w_i + \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik} + u_i$$
$$\mathrm{E}(u_i|w_i, x_{i1}, \ldots, x_{ik}) = 0.$$

What is the intercept ψ?

(iv) How would you estimate β (along with ψ and the γ_j) in part (iii)? Explain.

Computer Exercises

C1 A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u.$$

(i) What is the most likely sign for β_2?

(ii) Do you think *cigs* and *faminc* are likely to be correlated? Explain why the correlation might be positive or negative.

(iii) Now, estimate the equation with and without *faminc*, using the data in BWGHT. Report the results in equation form, including the sample size and R-squared. Discuss your results, focusing on whether adding *faminc* substantially changes the estimated effect of *cigs* on *bwght*.

C2 Use the data in HPRICE1 to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where *price* is the house price measured in thousands of dollars.

(i) Write out the results in equation form.

(ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

(iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

(iv) What percentage of the variation in price is explained by square footage and number of bedrooms?

(v) The first house in the sample has $sqrft = 2{,}438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

(vi) The actual selling price of the first house in the sample was \$300,000 (so $price = 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

C3 The file CEOSAL2 contains data on 177 chief executive officers and can be used to examine the effects of firm performance on CEO salary.

(i) Estimate a model relating annual salary to firm sales and market value. Make the model of the constant elasticity variety for both independent variables. Write the results out in equation form.

(ii) Add *profits* to the model from part (i). Why can this variable not be included in logarithmic form? Would you say that these firm performance variables explain most of the variation in CEO salaries?

(iii) Add the variable *ceoten* to the model in part (ii). What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?

(iv) Find the sample correlation coefficient between the variables log(*mktval*) and *profits*. Are these variables highly correlated? What does this say about the OLS estimators?

C4 Use the data in ATTEND for this exercise.

(i) Obtain the minimum, maximum, and average values for the variables *atndrte*, *priGPA*, and *ACT*.

(ii) Estimate the model

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u,$$

and write the results in equation form. Interpret the intercept. Does it have a useful meaning?

(iii) Discuss the estimated slope coefficients. Are there any surprises?

(iv) What is the predicted *atndrte* if $priGPA = 3.65$ and $ACT = 20$? What do you make of this result? Are there any students in the sample with these values of the explanatory variables?

(v) If Student A has $priGPA = 3.1$ and $ACT = 21$ and Student B has $priGPA = 2.1$ and $ACT = 26$, what is the predicted difference in their attendance rates?

C5 Confirm the partialling out interpretation of the OLS estimates by explicitly doing the partialling out for Example 3.2. This first requires regressing *educ* on *exper* and *tenure* and saving the residuals, $\hat{r}_1$. Then, regress log(*wage*) on $\hat{r}_1$. Compare the coefficient on $\hat{r}_1$ with the coefficient on *educ* in the regression of log(*wage*) on *educ*, *exper*, and *tenure*.

C6 Use the data set in WAGE2 for this problem. As usual, be sure all of the following regressions contain an intercept.

(i) Run a simple regression of *IQ* on *educ* to obtain the slope coefficient, say, $\tilde{\delta}_1$.

(ii) Run the simple regression of log(*wage*) on *educ,* and obtain the slope coefficient, $\tilde{\beta}_1$.

(iii) Run the multiple regression of log(*wage*) on *educ* and *IQ,* and obtain the slope coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

(iv) Verify that $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$.

C7 Use the data in MEAP93 to answer this question.

(i) Estimate the model

$$math10 = \beta_0 + \beta_1 \log(expend) + \beta_2 lnchprg + u,$$

and report the results in the usual form, including the sample size and *R*-squared. Are the signs of the slope coefficients what you expected? Explain.

(ii) What do you make of the intercept you estimated in part (i)? In particular, does it make sense to set the two explanatory variables to zero? [*Hint*: Recall that $\log(1) = 0$.]

(iii) Now run the simple regression of *math10* on log(*expend*), and compare the slope coefficient with the estimate obtained in part (i). Is the estimated spending effect now larger or smaller than in part (i)?

(iv) Find the correlation between $lexpend = \log(expend)$ and *lnchprg*. Does its sign make sense to you?

(v) Use part (iv) to explain your findings in part (iii).

C8 Use the data in DISCRIM to answer this question. These are ZIP code–level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

(i) Find the average values of *prpblck* and *income* in the sample, along with their standard deviations. What are the units of measurement of *prpblck* and *income*?

(ii) Consider a model to explain the price of soda, *psoda*, in terms of the proportion of the population that is black and median income:

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + u.$$

Estimate this model by OLS and report the results in equation form, including the sample size and *R*-squared. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on *prpblck*. Do you think it is economically large?

(iii) Compare the estimate from part (ii) with the simple regression estimate from *psoda* on *prpblck*. Is the discrimination effect larger or smaller when you control for income?

(iv) A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + u.$$

If *prpblck* increases by .20 (20 percentage points), what is the estimated percentage change in *psoda*? (*Hint*: The answer is 2.*xx*, where you fill in the "*xx*.")

(v) Now add the variable *prppov* to the regression in part (iv). What happens to $\hat{\beta}_{prpblck}$?

(vi) Find the correlation between log(*income*) and *prppov*. Is it roughly what you expected?

(vii) Evaluate the following statement: "Because log(*income*) and *prppov* are so highly correlated, they have no business being in the same regression."

C9 Use the data in CHARITY to answer the following questions:

(i) Estimate the equation

$$gift = \beta_0 + \beta_1 mailsyear + \beta_2 giftlast + \beta_3 propresp + u$$

by OLS and report the results in the usual way, including the sample size and *R*-squared. How does the *R*-squared compare with that from the simple regression that omits *giftlast* and *propresp*?

(ii) Interpret the coefficient on *mailsyear*. Is it bigger or smaller than the corresponding simple regression coefficient?

(iii) Interpret the coefficient on *propresp*. Be careful to notice the units of measurement of *propresp*.

(iv) Now add the variable *avggift* to the equation. What happens to the estimated effect of *mailsyear*?

(v) In the equation from part (iv), what has happened to the coefficient on *giftlast*? What do you think is happening?

C10 Use the data in HTV to answer this question. The data set includes information on wages, education, parents' education, and several other variables for 1,230 working men in 1991.

(i) What is the range of the *educ* variable in the sample? What percentage of men completed twelfth grade but no higher grade? Do the men or their parents have, on average, higher levels of education?

(ii) Estimate the regression model

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + u$$

by OLS and report the results in the usual form. How much sample variation in *educ* is explained by parents' education? Interpret the coefficient on *motheduc*.

(iii) Add the variable *abil* (a measure of cognitive ability) to the regression from part (ii), and report the results in equation form. Does "ability" help to explain variations in education, even after controlling for parents' education? Explain.

(iv) (Requires calculus) Now estimate an equation where *abil* appears in quadratic form:

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u.$$

Using the estimates $\hat{\beta}_3$ and $\hat{\beta}_4$, use calculus to find the value of *abil*, call it $abil^*$, where *educ* is minimized. (The other coefficients and values of parents' education variables have no effect; we are holding parents' education fixed.) Notice that *abil* is measured so that negative values are permissible. You might also verify that the second derivative is positive so that you do indeed have a minimum.

(v) Argue that only a small fraction of men in the sample have "ability" less than the value calculated in part (iv). Why is this important?

(vi) If you have access to a statistical program that includes graphing capabilities, use the estimates in part (iv) to graph the relationship between the predicted education and *abil*. Set *motheduc* and *fatheduc* at their average values in the sample, 12.18 and 12.45, respectively.

C11 Use the data in MEAPSINGLE to study the effects of single-parent households on student math performance. These data are for a subset of schools in southeast Michigan for the year 2000. The socioeconomic variables are obtained at the ZIP code level (where ZIP code is assigned to schools based on their mailing addresses).

(i) Run the simple regression of *math*4 on *pctsgle* and report the results in the usual format. Interpret the slope coefficient. Does the effect of single parenthood seem large or small?

(ii) Add the variables *lmedinc* and *free* to the equation. What happens to the coefficient on *pctsgle*? Explain what is happening.

(iii) Find the sample correlation between *lmedinc* and *free*. Does it have the sign you expect?

(iv) Does the substantial correlation between *lmedinc* and *free* mean that you should drop one from the regression to better estimate the causal effect of single parenthood on student performance? Explain.

(v) Find the variance inflation factors (VIFs) for each of the explanatory variables appearing in the regression in part (ii). Which variable has the largest VIF? Does this knowledge affect the model you would use to study the causal effect of single parenthood on math performance?

C12 The data in ECONMATH contain grade point averages and standardized test scores, along with performance in an introductory economics course, for students at a large public university. The variable to be explained is *score*, the final score in the course measured as a percentage.

(i) How many students received a perfect score for the course? What was the average score? Find the means and standard deviations of *actmth* and *acteng*, and discuss how they compare.

(ii) Estimate a linear equation relating *score* to *colgpa*, *actmth*, and *acteng*, where *colgpa* is measured at the beginning of the term. Report the results in the usual form.

(iii) Would you say the math or English ACT score is a better predictor of performance in the economics course? Explain.

(iv) Discuss the size of the *R*-squared in the regression.

C13 Use the data in GPA1 to answer this question. We can compare multiple regression estimates, where we control for student achievement and background variables, and compare our findings with the difference-in-means estimate in Computer Exercise C11 in Chapter 2.

(i) In the simple regression equation

$$colGPA = \beta_0 + \beta_1 PC + u$$

obtain $\hat{\beta}_0$ and $\hat{\beta}_1$. Interpret these estimates.

(ii) Now add the controls *hsGPA* and *ACT*—that is, run the regression *colGPA* on *PC, hsGPA, and ACT.* Does the coefficient on *PC* change much from part (ii)? Does $\hat{\beta}_{hsGPA}$ make sense?

(iii) In the estimation from part (ii), what is worth more: Owning a PC or having 10 more points on the ACT score?

(iv) Now to the regression in part (ii) add the two binary indicators for the parents being college graduates. Does the estimate of β_1 change much from part (ii)? How much variation are you explaining in *colGPA?*

(v) Suppose someone looking at your regression from part (iv) says to you, "The variables *hsGPA* and *ACT* are probably pretty highly correlated, so you should drop one of them from the regression." How would you respond?

APPENDIX 3A

3A.1 Derivation of the First Order Conditions in Equation (3.13)

The analysis is very similar to the simple regression case. We must characterize the solutions to the problem

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2.$$

Taking the partial derivatives with respect to each of the b_j (see Math Refresher A), evaluating them at the solutions, and setting them equal to zero gives

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$-2\sum_{i=1}^{n} x_{ij}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = 0, \quad \text{for all } j = 1, \ldots, k.$$

Canceling the -2 gives the first order conditions in (3.13).

3A.2 Derivation of Equation (3.22)

To derive (3.22), write x_{i1} in terms of its fitted value and its residual from the regression of x_1 on $x_2, \ldots, x_k$: $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$, for all $i = 1, \ldots, n$. Now, plug this into the second equation in (3.13):

$$\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = 0. \qquad [3.76]$$

By the definition of the OLS residual $\hat{u}_i$, because $\hat{x}_{i1}$ is just a linear function of the explanatory variables $x_{i2}, \ldots, x_{ik}$, it follows that $\sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i = 0$. Therefore, equation (3.76) can be expressed as

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = 0. \qquad [3.77]$$

Because the $\hat{r}_{i1}$ are the residuals from regressing x_1 on $x_2, \ldots, x_k$, $\sum_{i=1}^{n} x_{ij}\hat{r}_{i1} = 0$, for all $j = 2, \ldots, k$. Therefore, (3.77) is equivalent to $\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 x_{i1}) = 0$. Finally, we use the fact that $\sum_{i=1}^{n} \hat{x}_{i1}\hat{r}_{i1} = 0$, which means that $\hat{\beta}_1$ solves

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 \hat{r}_{i1}) = 0.$$

Now, straightforward algebra gives (3.22), provided, of course, that $\sum_{i=1}^{n} \hat{r}_{i1}^2 > 0$; this is ensured by Assumption MLR.3.

3A.3 Proof of Theorem 3.1

We prove Theorem 3.1 for $\hat{\beta}_1$; the proof for the other slope parameters is virtually identical. (See Advanced Treatment E for a more succinct proof using matrices.) Under Assumption MLR.3, the OLS estimators exist, and we can write $\hat{\beta}_1$ as in (3.22). Under Assumption MLR.1, we can write y_i as in (3.32); substitute this for y_i in (3.22). Then, using $\sum_{i=1}^{n} \hat{r}_{i1} = 0$, $\sum_{i=1}^{n} x_{ij}\hat{r}_{i1} = 0$, for all $j = 2, \ldots, k$, and $\sum_{i=1}^{n} x_{i1}\hat{r}_{i1} = \sum_{i=1}^{n} \hat{r}_{i1}^2$, we have

$$\hat{\beta}_1 = \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1} u_i\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right). \qquad [3.78]$$

Now, under Assumptions MLR.2 and MLR.4, the expected value of each u_i, given all independent variables in the sample, is zero. Because the $\hat{r}_{i1}$ are just functions of the sample independent variables, it follows that

$$\begin{aligned} \mathrm{E}(\hat{\beta}_1|\mathbf{X}) &= \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1}\mathrm{E}(u_i|\mathbf{X})\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) \\ &= \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1} \cdot 0\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) = \beta_1, \end{aligned}$$

where $\mathbf{X}$ denotes the data on all independent variables and $\mathrm{E}(\hat{\beta}_1|\mathbf{X})$ is the expected value of $\hat{\beta}_1$, given $x_{i1}, \ldots, x_{ik}$, for all $i = 1, \ldots, n$. This completes the proof.

3A.4 General Omitted Variable Bias

We can derive the omitted variable bias in the general model in equation (3.31) under the first four Gauss-Markov assumptions. In particular, let the $\hat{\beta}_j$, $j = 0, 1, \ldots, k$ be the OLS estimators from the regression using the full set of explanatory variables. Let the $\tilde{\beta}_j$, $j = 0, 1, \ldots, k-1$ be the OLS estimators from the regression that leaves out x_k. Let $\tilde{\delta}_j$, $j = 1, \ldots, k-1$ be the slope coefficient on x_j in the auxiliary regression of x_{ik} on $x_{i1}, x_{i2}, \ldots x_{i,k-1}$, $i = 1, \ldots, n$. A useful fact is that

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k \tilde{\delta}_j. \qquad [3.79]$$

This shows explicitly that, when we do not control for x_k in the regression, the estimated partial effect of x_j equals the partial effect when we include x_k plus the partial effect of x_k on $\hat{y}$ times the partial relationship between the omitted variable, x_k, and x_j, $j < k$. Conditional on the entire set of explanatory variables, $\mathbf{X}$, we know that the $\hat{\beta}_j$ are all unbiased for the corresponding β_j, $j = 1, \ldots, k$. Further, because $\tilde{\delta}_j$ is just a function of $\mathbf{X}$, we have

$$\begin{aligned} \mathrm{E}(\tilde{\beta}_j|\mathbf{X}) &= \mathrm{E}(\hat{\beta}_j|\mathbf{X}) + \mathrm{E}(\hat{\beta}_k|\mathbf{X})\tilde{\delta}_j \\ &= \beta_j + \beta_k \tilde{\delta}_j. \end{aligned} \qquad [3.80]$$

Equation (3.80) shows that $\tilde{\beta}_j$ is biased for β_j unless $\beta_k = 0$—in which case x_k has no partial effect in the population—or $\tilde{\delta}_j$ equals zero, which means that x_{ik} and x_{ij} are partially uncorrelated in the sample. The key to obtaining equation (3.80) is equation (3.79). To show equation (3.79), we can use equation (3.22) a couple of times. For simplicity, we look at $j = 1$. Now, $\tilde{\beta}_1$ is the slope coefficient in the simple regression of y_i on $\tilde{r}_{i1}$, $i = 1, \ldots, n$, where the $\tilde{r}_{i1}$ are the OLS residuals from the regression of x_{i1} on $x_{i2}, x_{i3}, \ldots, x_{i,k-1}$. Consider the numerator of the expression for $\tilde{\beta}_1$: $\sum_{i=1}^{n} \tilde{r}_{i1} y_i$. But for each i, we can write $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} + \hat{u}_i$ and plug in for y_i. Now, by properties of the OLS residuals, the $\tilde{r}_{i1}$ have zero sample average and are uncorrelated with $x_{i2}, x_{i3}, \ldots, x_{i,k-1}$ in the sample. Similarly, the $\hat{u}_i$ have zero sample average and zero sample correlation with $x_{i1}, x_{i2}, \ldots, x_{ik}$. It follows that the $\tilde{r}_{i1}$ and $\hat{u}_i$ are uncorrelated in the sample (because the $\tilde{r}_{i1}$ are just linear combinations of $x_{i1}, x_{i2}, \ldots, x_{i,k-1}$). So

$$\sum_{i=1}^{n} \tilde{r}_{i1} y_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} \right) + \hat{\beta}_k \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} \right). \qquad [3.81]$$

Now, $\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} = \sum_{i=1}^{n} \tilde{r}_{i1}^2$, which is also the denominator of $\tilde{\beta}_1$. Therefore, we have shown that

$$\begin{aligned} \tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_k \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} \right) \Big/ \left(\sum_{i=1}^{n} \tilde{r}_{i1}^2 \right) \\ &= \hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1. \end{aligned}$$

This is the relationship we wanted to show.

3A.5 Proof of Theorem 3.2

Again, we prove this for $j = 1$. Write $\hat{\beta}_1$ as in equation (3.78). Now, under MLR.5, $\text{Var}(u_i|\mathbf{X}) = \sigma^2$, for all $i = 1, \ldots, n$. Under random sampling, the u_i are independent, even conditional on $\mathbf{X}$, and the $\hat{r}_{i1}$ are nonrandom conditional on $\mathbf{X}$. Therefore,

$$\begin{aligned} \text{Var}(\hat{\beta}_1|\mathbf{X}) &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \, \text{Var}(u_i|\mathbf{X}) \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)^2 \\ &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \sigma^2 \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)^2 = \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \end{aligned}$$

Now, because $\sum_{i=1}^{n} \hat{r}_{i1}^2$ is the sum of squared residuals from regressing x_1 on $x_2, \ldots, x_k$, $\sum_{i=1}^{n} \hat{r}_{i1}^2 = \text{SST}_1(1 - R_1^2)$. This completes the proof.

3A.6 Proof of Theorem 3.4

We show that, for any other linear unbiased estimator $\tilde{\beta}_1$ of β_1, $\text{Var}(\tilde{\beta}_1) \geq \text{Var}(\hat{\beta}_1)$, where $\hat{\beta}_1$ is the OLS estimator. The focus on $j = 1$ is without loss of generality.

For $\tilde{\beta}_1$ as in equation (3.60), we can plug in for y_i to obtain

$$\tilde{\beta}_1 = \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1} x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} + \sum_{i=1}^{n} w_{i1} u_i.$$

Now, because the w_{i1} are functions of the x_{ij},

$$\begin{aligned} \text{E}(\tilde{\beta}_1|\mathbf{X}) &= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1} x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} + \sum_{i=1}^{n} w_{i1} E(u_i|\mathbf{X}) \\ &= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1} x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} \end{aligned}$$

because $E(u_i|\mathbf{X}) = 0$, for all $i = 1, \ldots, n$ under MLR.2 and MLR.4. Therefore, for $E(\tilde{\beta}_1|\mathbf{X})$ to equal β_1 for any values of the parameters, we must have

$$\sum_{i=1}^{n} w_{i1} = 0, \sum_{i=1}^{n} w_{i1}x_{i1} = 1, \sum_{i=1}^{n} w_{i1}x_{ij} = 0, \; j = 2, \ldots, k. \qquad [3.82]$$

Now, let $\hat{r}_{i1}$ be the residuals from the regression of x_{i1} on $x_{i2}, \ldots, x_{ik}$. Then, from (3.82), it follows that

$$\sum_{i=1}^{n} w_{i1}r_{i1} = 1 \qquad [3.83]$$

because $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$ and $\sum_{i=1}^{n} w_{i1}\hat{x}_{i1} = 0$. Now, consider the difference between $\text{Var}(\tilde{\beta}_1|\mathbf{X})$ and $\text{Var}(\hat{\beta}_1|\mathbf{X})$ under MLR.1 through MLR.5:

$$\sigma^2 \sum_{i=1}^{n} w_{i1}^2 - \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \qquad [3.84]$$

Because of (3.83), we can write the difference in (3.84), without σ^2, as

$$\sum_{i=1}^{n} w_{i1}^2 - \left(\sum_{i=1}^{n} w_{i1}\hat{r}_{i1} \right)^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \qquad [3.85]$$

But (3.85) is simply

$$\sum_{i=1}^{n} (w_{i1} - \hat{\gamma}_1\hat{r}_{i1})^2, \qquad [3.86]$$

where $\hat{\gamma}_1 = (\sum_{i=1}^{n} w_{i1}\hat{r}_{i1})/(\sum_{i=1}^{n} \hat{r}_{i1}^2)$, as can be seen by squaring each term in (3.86), summing, and then canceling terms. Because (3.86) is just the sum of squared residuals from the simple regression of w_{i1} on $\hat{r}_{i1}$—remember that the sample average of $\hat{r}_{i1}$ is zero—(3.86) must be nonnegative. This completes the proof.

CHAPTER 4

Multiple Regression Analysis: Inference

This chapter continues our treatment of multiple regression analysis. We now turn to the problem of testing hypotheses about the parameters in the population regression model. We begin in Section 4-1 by finding the distributions of the OLS estimators under the added assumption that the population error is normally distributed. Sections 4-2 and 4-3 cover hypothesis testing about individual parameters, while Section 4-4 discusses how to test a single hypothesis involving more than one parameter. We focus on testing multiple restrictions in Section 4-5 and pay particular attention to determining whether a group of independent variables can be omitted from a model.

4-1 Sampling Distributions of the OLS Estimators

Up to this point, we have formed a set of assumptions under which OLS is unbiased; we have also derived and discussed the bias caused by omitted variables. In Section 3-4, we obtained the variances of the OLS estimators under the Gauss-Markov assumptions. In Section 3-5, we showed that this variance is smallest among linear unbiased estimators.

Knowing the expected value and variance of the OLS estimators is useful for describing the precision of the OLS estimators. However, in order to perform statistical inference, we need to know more than just the first two moments of $\hat{\beta}_j$; we need to know the full sampling distribution of the $\hat{\beta}_j$. Even under the Gauss-Markov assumptions, the distribution of $\hat{\beta}_j$ can have virtually any shape.

When we condition on the values of the independent variables in our sample, it is clear that the sampling distributions of the OLS estimators depend on the underlying distribution of the errors. To make the sampling distributions of the $\hat{\beta}_j$ tractable, we now assume that the unobserved error is *normally distributed* in the population. We call this the **normality assumption**.

Assumption MLR.6 **Normality**

The population error u is *independent* of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0,\sigma^2)$.

Assumption MLR.6 is much stronger than any of our previous assumptions. In fact, because u is independent of the x_j under MLR.6, $\text{E}(u|x_1, \ldots, x_k) = \text{E}(u) = 0$ and $\text{Var}(u|x_1, \ldots, x_k) = \text{Var}(u) = \sigma^2$. Thus, if we make Assumption MLR.6, then we are necessarily assuming MLR.4 and MLR.5. To emphasize that we are assuming more than before, we will refer to the full set of Assumptions MLR.1 through MLR.6.

For cross-sectional regression applications, Assumptions MLR.1 through MLR.6 are called the **classical linear model (CLM) assumptions**. Thus, we will refer to the model under these six assumptions as the **classical linear model**. It is best to think of the CLM assumptions as containing all of the Gauss-Markov assumptions *plus* the assumption of a normally distributed error term.

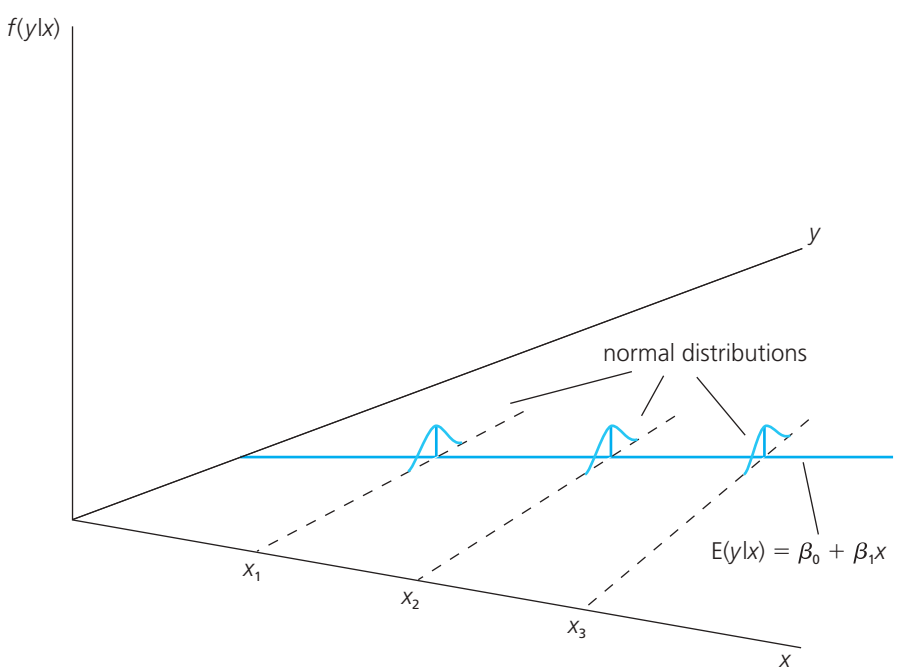
Under the CLM assumptions, the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ have a stronger efficiency property than they would under the Gauss-Markov assumptions. It can be shown that the OLS estimators are the **minimum variance unbiased estimators**, which means that OLS has the smallest variance among unbiased estimators; we no longer have to restrict our comparison to estimators that are linear in the y_i. This property of OLS under the CLM assumptions is discussed further in Advanced Treatment E.

A succinct way to summarize the population assumptions of the CLM is

$$y|\mathbf{x} \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \sigma^2),$$

where $\mathbf{x}$ is again shorthand for $(x_1, \ldots, x_k)$. Thus, conditional on $\mathbf{x}$, y has a normal distribution with mean linear in $x_1, \ldots, x_k$ and a constant variance. For a single independent variable x, this situation is shown in Figure 4.1.

Figure 4.1 The homoskedastic normal distribution with a single explanatory variable.



The argument justifying the normal distribution for the errors usually runs something like this: Because u is the sum of many different unobserved factors affecting y, we can invoke the central limit theorem (CLT) (see Math Refresher C) to conclude that u has an approximate normal distribution. This argument has some merit, but it is not without weaknesses. First, the factors in u can have very different distributions in the population (for example, ability and quality of schooling in the error in a wage equation). Although the CLT can still hold in such cases, the normal approximation can be poor depending on how many factors appear in u and how different their distributions are.

A more serious problem with the CLT argument is that it assumes that all unobserved factors affect y in a separate, additive fashion. Nothing guarantees that this is so. If u is a complicated function of the unobserved factors, then the CLT argument does not really apply.

In any application, whether normality of u can be assumed is really an empirical matter. For example, there is no theorem that says *wage* conditional on *educ*, *exper*, and *tenure* is normally distributed. If anything, simple reasoning suggests that the opposite is true: because *wage* can never be less than zero, it cannot, strictly speaking, have a normal distribution. Further, because there are minimum wage laws, some fraction of the population earns exactly the minimum wage, which also violates the normality assumption. Nevertheless, as a practical matter, we can ask whether the conditional wage distribution is "close" to being normal. Past empirical evidence suggests that normality is *not* a good assumption for wages.

Often, using a transformation, especially taking the log, yields a distribution that is closer to normal. For example, something like log(*price*) tends to have a distribution that looks more normal than the distribution of *price*. Again, this is an empirical issue. We will discuss the consequences of nonnormality for statistical inference in Chapter 5.

There are some applications where MLR.6 is clearly false, as can be demonstrated with simple introspection. Whenever y takes on just a few values it cannot have anything close to a normal distribution. The dependent variable in Example 3.5 provides a good example. The variable *narr86*, the number of times a young man was arrested in 1986, takes on a small range of integer values and is zero for most men. Thus, *narr86* is far from being normally distributed. What can be done in these cases? As we will see in Chapter 5—and this is important—nonnormality of the errors is not a serious problem with large sample sizes. For now, we just make the normality assumption.

Normality of the error term translates into normal sampling distributions of the OLS estimators:

THEOREM 4.1

NORMAL SAMPLING DISTRIBUTIONS

Under the CLM assumptions MLR.1 through MLR.6, conditional on the sample values of the independent variables,

$$\hat{\beta}_j \sim \text{Normal}[\beta_j, \text{Var}(\hat{\beta}_j)], \qquad [4.1]$$

where $\text{Var}(\hat{\beta}_j)$ was given in Chapter 3 [equation (3.51)]. Therefore,

$$(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \sim \text{Normal}(0,1).$$

The proof of (4.1) is not that difficult, given the properties of normally distributed random variables in Math Refresher B. Each $\hat{\beta}_j$ can be written as $\hat{\beta}_j = \beta_j + \sum_{i=1}^{n} w_{ij} u_i$, where $w_{ij} = \hat{r}_{ij}/\text{SSR}_j$, $\hat{r}_{ij}$ is the i^{th} residual from the regression of the x_j on all the other independent variables, and SSR_j is the sum of squared residuals from this regression [see equation (3.65)]. Because the w_{ij} depend only on the independent variables, they can be treated as nonrandom. Thus, $\hat{\beta}_j$ is just a linear combination of the errors in the sample $\{u_i: i = 1, 2, \ldots, n\}$. Under Assumption MLR.6 (and the random sampling Assumption MLR.2), the errors are independent, identically distributed $\text{Normal}(0,\sigma^2)$ random variables. An important fact about independent normal random variables is that a linear combination of such random variables is normally

GOING FURTHER 4.1

Suppose that u is independent of the explanatory variables, and it takes on the values -2, -1, 0, 1, and 2 with equal probability of 1/5. Does this violate the Gauss-Markov assumptions? Does this violate the CLM assumptions?

distributed (see Math Refresher B). This basically completes the proof. In Section 3-3, we showed that $E(\hat{\beta}_j) = \beta_j$, and we derived $Var(\hat{\beta}_j)$ in Section 3-4; there is no need to re-derive these facts.

The second part of this theorem follows immediately from the fact that when we standardize a normal random variable by subtracting off its mean and dividing by its standard deviation, we end up with a standard normal random variable.

The conclusions of Theorem 4.1 can be strengthened. In addition to (4.1), any linear combination of the $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ is also normally distributed, and any subset of the $\hat{\beta}_j$ has a *joint* normal distribution. These facts underlie the testing results in the remainder of this chapter. In Chapter 5, we will show that the normality of the OLS estimators is still *approximately* true in large samples even without normality of the errors.

4-2 Testing Hypotheses about a Single Population Parameter: The *t* Test

This section covers the very important topic of testing hypotheses about any single parameter in the population regression function. The population model can be written as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \qquad [4.2]$$

and we assume that it satisfies the CLM assumptions. We know that OLS produces unbiased estimators of the β_j. In this section, we study how to test hypotheses about a particular β_j. For a full understanding of hypothesis testing, one must remember that the β_j are unknown features of the population, and we will never know them with certainty. Nevertheless, we can *hypothesize* about the value of β_j and then use statistical inference to test our hypothesis.

In order to construct hypotheses tests, we need the following result:

THEOREM 4.2

***t* DISTRIBUTION FOR THE STANDARDIZED ESTIMATORS**

Under the CLM assumptions MLR.1 through MLR.6,

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \sim t_{n-k-1} = t_{df}, \qquad [4.3]$$

where $k + 1$ is the number of unknown parameters in the population model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ (k slope parameters and the intercept β_0) and $n - k - 1$ is the degrees of freedom (df).

This result differs from Theorem 4.1 in some notable respects. Theorem 4.1 showed that, under the CLM assumptions, $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \sim \text{Normal}(0,1)$. The t distribution in (4.3) comes from the fact that the constant σ in $\text{sd}(\hat{\beta}_j)$ has been replaced with the random variable $\hat{\sigma}$. The proof that this leads to a t distribution with $n - k - 1$ degrees of freedom is difficult and not especially instructive. Essentially, the proof shows that (4.3) can be written as the ratio of the standard normal random variable $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j)$ over the square root of $\hat{\sigma}^2/\sigma^2$. These random variables can be shown to be independent, and $(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}$. The result then follows from the definition of a t random variable (see Section B-5 in Math Refresher B).

Theorem 4.2 is important in that it allows us to test hypotheses involving the β_j. In most applications, our primary interest lies in testing the **null hypothesis**

$$H_0: \beta_j = 0, \qquad [4.4]$$

where j corresponds to any of the k independent variables. It is important to understand what (4.4) means and to be able to describe this hypothesis in simple language for a particular application. Because β_j measures the partial effect of x_j on (the expected value of) y, after controlling for all other independent variables, (4.4) means that, once $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ have been accounted for, x_j has *no effect* on the expected value of y. We cannot state the null hypothesis as "x_j does have a partial effect on y" because this is true for any value of β_j other than zero. Classical testing is suited for testing *simple hypotheses* like (4.4).

As an example, consider the wage equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

The null hypothesis H_0: $\beta_2 = 0$ means that, once education and tenure have been accounted for, the number of years in the workforce (*exper*) has no effect on hourly wage. This is an economically interesting hypothesis. If it is true, it implies that a person's work history prior to the current employment does not affect wage. If $\beta_2 > 0$, then prior work experience contributes to productivity, and hence to wage.

You probably remember from your statistics course the rudiments of hypothesis testing for the mean from a normal population. (This is reviewed in Math Refresher C.) The mechanics of testing (4.4) in the multiple regression context are very similar. The hard part is obtaining the coefficient estimates, the standard errors, and the critical values, but most of this work is done automatically by econometrics software. Our job is to learn how regression output can be used to test hypotheses of interest.

The statistic we use to test (4.4) (against any alternative) is called "the" ***t* statistic** or "the" ***t* ratio** of $\hat{\beta}_j$ and is defined as

$$t_{\hat{\beta}_j} \equiv \hat{\beta}_j / \text{se}(\hat{\beta}_j). \qquad [4.5]$$

We have put "the" in quotation marks because, as we will see shortly, a more general form of the t statistic is needed for testing other hypotheses about β_j. For now, it is important to know that (4.5) is suitable only for testing (4.4). For particular applications, it is helpful to index t statistics using the name of the independent variable; for example, t_{educ} would be the t statistic for $\hat{\beta}_{educ}$.

The t statistic for $\hat{\beta}_j$ is simple to compute given $\hat{\beta}_j$ and its standard error. In fact, most regression packages do the division for you and report the t statistic along with each coefficient and its standard error.

Before discussing how to use (4.5) formally to test H_0: $\beta_j = 0$, it is useful to see why $t_{\hat{\beta}_j}$ has features that make it reasonable as a test statistic to detect $\beta_j \neq 0$. First, because $\text{se}(\hat{\beta}_j)$ is always positive, $t_{\hat{\beta}_j}$ has the same sign as $\hat{\beta}_j$: if $\hat{\beta}_j$ is positive, then so is $t_{\hat{\beta}_j}$ and if $\hat{\beta}_j$ is negative, so is $t_{\hat{\beta}_j}$. Second, for a given value of $\text{se}(\hat{\beta}_j)$, a larger value of $\hat{\beta}_j$ leads to larger values of $t_{\hat{\beta}_j}$. If $\hat{\beta}_j$ becomes more negative, so does $t_{\hat{\beta}_j}$.

Because we are testing H_0: $\beta_j = 0$, it is only natural to look at our unbiased estimator of β_j, $\hat{\beta}_j$, for guidance. In any interesting application, the point estimate $\hat{\beta}_j$ will *never* exactly be zero, whether or not H_0 is true. The question is: How far is $\hat{\beta}_j$ from zero? A sample value of $\hat{\beta}_j$ very far from zero provides evidence against H_0: $\beta_j = 0$. However, we must recognize that there is a sampling error in our estimate $\hat{\beta}_j$, so the size of $\hat{\beta}_j$ must be weighed against its sampling error. Because the standard error of $\hat{\beta}_j$ is an estimate of the standard deviation of $\hat{\beta}_j$, $t_{\hat{\beta}_j}$ measures how many estimated standard deviations $\hat{\beta}_j$ is away from zero. This is precisely what we do in testing whether the mean of a population is zero, using the standard t statistic from introductory statistics. Values of $t_{\hat{\beta}_j}$ sufficiently far from zero will result in a rejection of H_0. The precise rejection rule depends on the alternative hypothesis and the chosen significance level of the test.

Determining a rule for rejecting (4.4) at a given significance level—that is, the probability of rejecting H_0 when it is true—requires knowing the sampling distribution of $t_{\hat{\beta}_j}$ when H_0 is true. From Theorem 4.2, we know this to be t_{n-k-1}. This is the key theoretical result needed for testing (4.4).

Before proceeding, it is important to remember that we are testing hypotheses about the *population* parameters. We are *not* testing hypotheses about the estimates from a particular sample. Thus, it

never makes sense to state a null hypothesis as "H_0: $\hat{\beta}_1 = 0$" or, even worse, as "H_0: .237 = 0" when the estimate of a parameter is .237 in the sample. We are testing whether the unknown population value, β_1, is zero.

Some treatments of regression analysis define the *t* statistic as the *absolute value* of (4.5), so that the *t* statistic is always positive. This practice has the drawback of making testing against one-sided alternatives clumsy. Throughout this text, the *t* statistic always has the same sign as the corresponding OLS coefficient estimate.

4-2a Testing against One-Sided Alternatives

To determine a rule for rejecting H_0, we need to decide on the relevant **alternative hypothesis**. First, consider a **one-sided alternative** of the form

$$H_1: \beta_j > 0. \qquad [4.6]$$

When we state the alternative as in equation (4.6), we are really saying that the null hypothesis is H_0: $\beta_j \leq 0$. For example, if β_j is the coefficient on education in a wage regression, we only care about detecting that β_j is different from zero when β_j is actually positive. You may remember from introductory statistics that the null value that is hardest to reject in favor of (4.6) is $\beta_j = 0$. In other words, if we reject the null $\beta_j = 0$ then we automatically reject $\beta_j < 0$. Therefore, it suffices to act as if we are testing H_0: $\beta_j = 0$ against H_1: $\beta_j > 0$, effectively ignoring $\beta_j < 0$, and that is the approach we take in this book.

How should we choose a rejection rule? We must first decide on a **significance level** ("level" for short) or the probability of rejecting H_0 when it is in fact true. For concreteness, suppose we have decided on a 5% significance level, as this is the most popular choice. Thus, we are willing to mistakenly reject H_0 when it is true 5% of the time. Now, while $t_{\hat{\beta}_j}$ has a *t* distribution under H_0—so that it has zero mean—under the alternative $\beta_j > 0$, the expected value of $t_{\hat{\beta}_j}$ is positive. Thus, we are looking for a "sufficiently large" positive value of $t_{\hat{\beta}_j}$ in order to reject H_0: $\beta_j = 0$ in favor of H_1: $\beta_j > 0$. Negative values of $t_{\hat{\beta}_j}$ provide no evidence in favor of H_1.

The definition of "sufficiently large," with a 5% significance level, is the 95th percentile in a *t* distribution with $n - k - 1$ degrees of freedom; denote this by *c*. In other words, the **rejection rule** is that H_0 is rejected in favor of H_1 at the 5% significance level if

$$t_{\hat{\beta}_j} > c. \qquad [4.7]$$

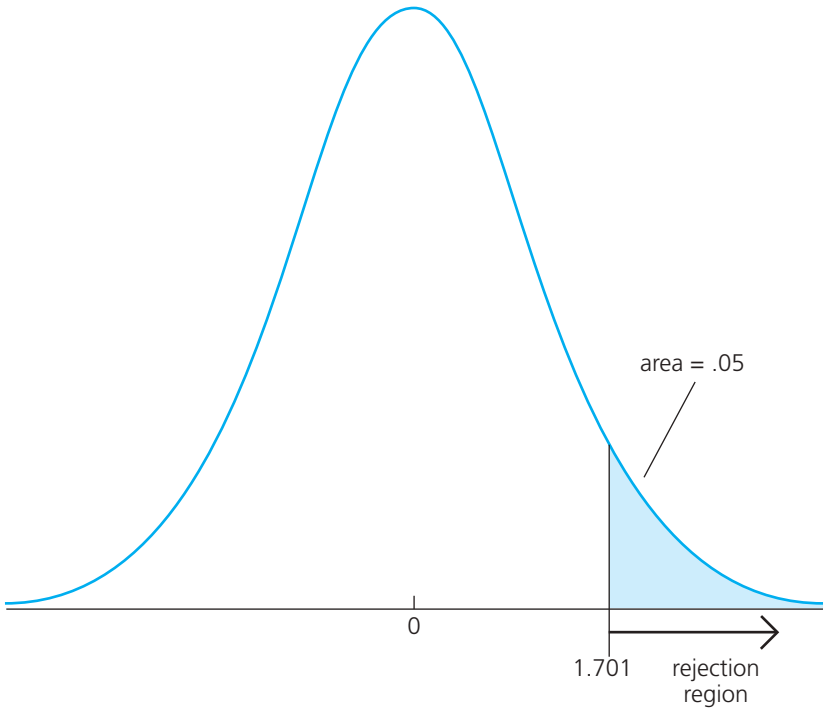
By our choice of the **critical value**, *c*, rejection of H_0 will occur for 5% of all random samples when H_0 is true.

The rejection rule in (4.7) is an example of a **one-tailed test**. To obtain *c*, we only need the significance level and the degrees of freedom. For example, for a 5% level test and with $n - k - 1 = 28$ degrees of freedom, the critical value is $c = 1.701$. If $t_{\hat{\beta}_j} \leq 1.701$, then we fail to reject H_0 in favor of (4.6) at the 5% level. Note that a negative value for $t_{\hat{\beta}_j}$, no matter how large in absolute value, leads to a failure in rejecting H_0 in favor of (4.6). (See Figure 4.2.)

The same procedure can be used with other significance levels. For a 10% level test and if $df = 21$, the critical value is $c = 1.323$. For a 1% significance level and if $df = 21$, $c = 2.518$. All of these critical values are obtained directly from Table G.2. You should note a pattern in the critical values: as the significance level falls, the critical value increases, so that we require a larger and larger value of $t_{\hat{\beta}_j}$ in order to reject H_0. Thus, if H_0 is rejected at, say, the 5% level, then it is automatically rejected at the 10% level as well. It makes no sense to reject the null hypothesis at, say, the 5% level and then to redo the test to determine the outcome at the 10% level.

As the degrees of freedom in the *t* distribution get large, the *t* distribution approaches the standard normal distribution. For example, when $n - k - 1 = 120$, the 5% critical value for the one-sided alternative (4.7) is 1.658, compared with the standard normal value of 1.645. These are close enough for practical purposes; for degrees of freedom greater than 120, one can use the standard normal critical values.

Figure 4.2 5% rejection rule for the alternative H_1: $\beta_j > 0$ with 28 *df*.



EXAMPLE 4.1 **Hourly Wage Equation**

Using the data in WAGE1 gives the estimated equation

$$\widehat{\log(wage)} = .284 + .092\,educ + .0041\,exper + .022\,tenure$$
$$(.104)\ (.007)\qquad (.0017)\qquad (.003)$$
$$n = 526,\ R^2 = .316,$$

where standard errors appear in parentheses below the estimated coefficients. We will follow this convention throughout the text. This equation can be used to test whether the return to *exper*, controlling for *educ* and *tenure*, is zero in the population, against the alternative that it is positive. Write this as H_0: $\beta_{exper} = 0$ versus H_1: $\beta_{exper} > 0$. (In applications, indexing a parameter by its associated variable name is a nice way to label parameters, because the numerical indices that we use in the general model are arbitrary and can cause confusion.) Remember that β_{exper} denotes the unknown population parameter. It is nonsense to write "H_0: .0041 = 0" or "H_0: $\hat{\beta}_{exper} = 0$."

Because we have 522 degrees of freedom, we can use the standard normal critical values. The 5% critical value is 1.645, and the 1% critical value is 2.326. The *t* statistic for $\hat{\beta}_{exper}$ is

$$t_{exper} = .0041/.0017 \approx 2.41,$$

and so $\hat{\beta}_{exper}$, or *exper*, is statistically significant even at the 1% level. We also say that "$\hat{\beta}_{exper}$ is statistically greater than zero at the 1% significance level."

The estimated return for another year of experience, holding tenure and education fixed, is not especially large. For example, adding three more years increases log(*wage*) by $3(.0041) = .0123$, so wage is only about 1.2% higher. Nevertheless, we have persuasively shown that the partial effect of experience *is* positive in the population.

The one-sided alternative that the parameter is less than zero,

$$H_1: \beta_j < 0, \qquad [4.8]$$

also arises in applications. The rejection rule for alternative (4.8) is just the mirror image of the previous case. Now, the critical value comes from the left tail of the t distribution. In practice, it is easiest to think of the rejection rule as

$$t_{\hat{\beta}_j} < -c, \qquad [4.9]$$

where c is the critical value for the alternative $H_1: \beta_j > 0$. For simplicity, we always assume c is positive, because this is how critical values are reported in t tables, and so the critical value $-c$ is a negative number.

For example, if the significance level is 5% and the degrees of freedom is 18, then $c = 1.734$, and so $H_0: \beta_j = 0$ is rejected in favor of $H_1: \beta_j < 0$ at the 5% level if $t_{\hat{\beta}_j} < -1.734$. It is important to remember that, to reject H_0 against the negative alternative (4.8), we must get a negative t statistic. A positive t ratio, no matter how large, provides no evidence in favor of (4.8). The rejection rule is illustrated in Figure 4.3.

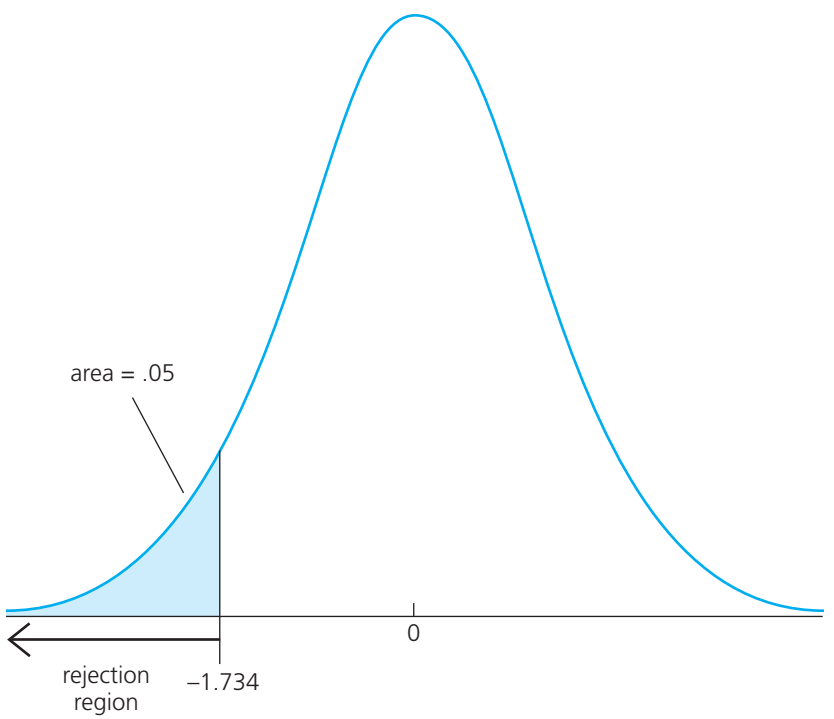
GOING FURTHER 4.2

Let community loan approval rates be determined by

$$apprate = \beta_0 + \beta_1 percmin + \beta_2 avginc + \beta_3 avgwlth + \beta_4 avgdebt + u,$$

where *percmin* is the percentage minority in the community, *avginc* is average income, *avgwlth* is average wealth, and *avgdebt* is some measure of average debt obligations. How do you state the null hypothesis that there is *no* difference in loan rates across neighborhoods due to racial and ethnic composition, when average income, average wealth, and average debt have been controlled for? How do you state the alternative that there is discrimination against minorities in loan approval rates?

Figure 4.3 5% rejection rule for the alternative $H_1: \beta_j < 0$ with 18 *df*.



EXAMPLE 4.2 **Student Performance and School Size**

There is much interest in the effect of school size on student performance. (See, for example, *The New York Times Magazine*, 5/28/95.) One claim is that, everything else being equal, students at smaller schools fare better than those at larger schools. This hypothesis is assumed to be true even after accounting for differences in class sizes across schools.

The file MEAP93 contains data on 408 high schools in Michigan for the year 1993. We can use these data to test the null hypothesis that school size has no effect on standardized test scores against the alternative that size has a negative effect. Performance is measured by the percentage of students receiving a passing score on the Michigan Educational Assessment Program (MEAP) standardized tenth-grade math test (*math*10). School size is measured by student enrollment (*enroll*). The null hypothesis is H_0: $\beta_{enroll} = 0$, and the alternative is H_1: $\beta_{enroll} < 0$. For now, we will control for two other factors, average annual teacher compensation (*totcomp*) and the number of staff per one thousand students (*staff*). Teacher compensation is a measure of teacher quality, and staff size is a rough measure of how much attention students receive.

The estimated equation, with standard errors in parentheses, is

$$\widehat{math10} = 2.274 + .00046\,totcomp + .048\,staff - .00020\,enroll$$
$$\qquad\qquad (6.113)\ (.00010)\qquad\quad (.040)\qquad (.00022)$$
$$n = 408,\ R^2 = .0541.$$

The coefficient on *enroll*, $-.00020$, is in accordance with the conjecture that larger schools hamper performance: higher enrollment leads to a lower percentage of students with a passing tenth-grade math score. (The coefficients on *totcomp* and *staff* also have the signs we expect.) The fact that *enroll* has an estimated coefficient different from zero could just be due to sampling error; to be convinced of an effect, we need to conduct a t test.

Because $n - k - 1 = 408 - 4 = 404$, we use the standard normal critical value. At the 5% level, the critical value is -1.65; the t statistic on *enroll* must be *less* than -1.65 to reject H_0 at the 5% level.

The t statistic on *enroll* is $-.00020/.00022 \approx -.91$, which is larger than -1.65: we *fail* to reject H_0 in favor of H_1 at the 5% level. In fact, the 15% critical value is -1.04, and because $-.91 > -1.04$, we fail to reject H_0 even at the 15% level. We conclude that *enroll* is not statistically significant at the 15% level.

The variable *totcomp* is statistically significant even at the 1% significance level because its t statistic is 4.6. On the other hand, the t statistic for *staff* is 1.2, and so we cannot reject H_0: $\beta_{staff} = 0$ against H_1: $\beta_{staff} > 0$ even at the 10% significance level. (The critical value is $c = 1.28$ from the standard normal distribution.)

To illustrate how changing functional form can affect our conclusions, we also estimate the model with all independent variables in logarithmic form. This allows, for example, the school size effect to diminish as school size increases. The estimated equation is

$$\widehat{math10} = -207.66 + 21.16\log(totcomp) + 3.98\log(staff) - 1.29\log(enroll)$$
$$\qquad\qquad (48.70)\ \ (4.06)\qquad\qquad (4.19)\qquad\quad (0.69)$$
$$n = 408,\ R^2 = .0654.$$

The t statistic on $\log(enroll)$ is about -1.87; because this is below the 5% critical value -1.65, we reject H_0: $\beta_{\log(enroll)} = 0$ in favor of H_1: $\beta_{\log(enroll)} < 0$ at the 5% level.

In Chapter 2, we encountered a model in which the dependent variable appeared in its original form (called *level* form), while the independent variable appeared in log form (called *level-log* model). The interpretation of the parameters is the same in the multiple regression context, except, of course, that we can give the parameters a ceteris paribus interpretation. Holding *totcomp* and *staff* fixed, we have $\Delta\widehat{math10} = -1.29[\Delta\log(enroll)]$, so that

$$\Delta\widehat{math10} \approx -(1.29/100)(\%\Delta enroll) \approx -.013(\%\Delta enroll).$$

Once again, we have used the fact that the change in log(*enroll*), when multiplied by 100, is approximately the percentage change in *enroll*. Thus, if enrollment is 10% higher at a school, $\widehat{math10}$ is predicted to be $.013(10) = 0.13$ percentage points lower (*math*10 is measured as a percentage).

Which model do we prefer, the one using the level of *enroll* or the one using log(*enroll*)? In the level-level model, enrollment does not have a statistically significant effect, but in the level-log model it does. This translates into a higher *R*-squared for the level-log model, which means we explain more of the variation in *math*10 by using *enroll* in logarithmic form (6.5% to 5.4%). The level-log model is preferred because it more closely captures the relationship between *math*10 and *enroll*. We will say more about using *R*-squared to choose functional form in Chapter 6.

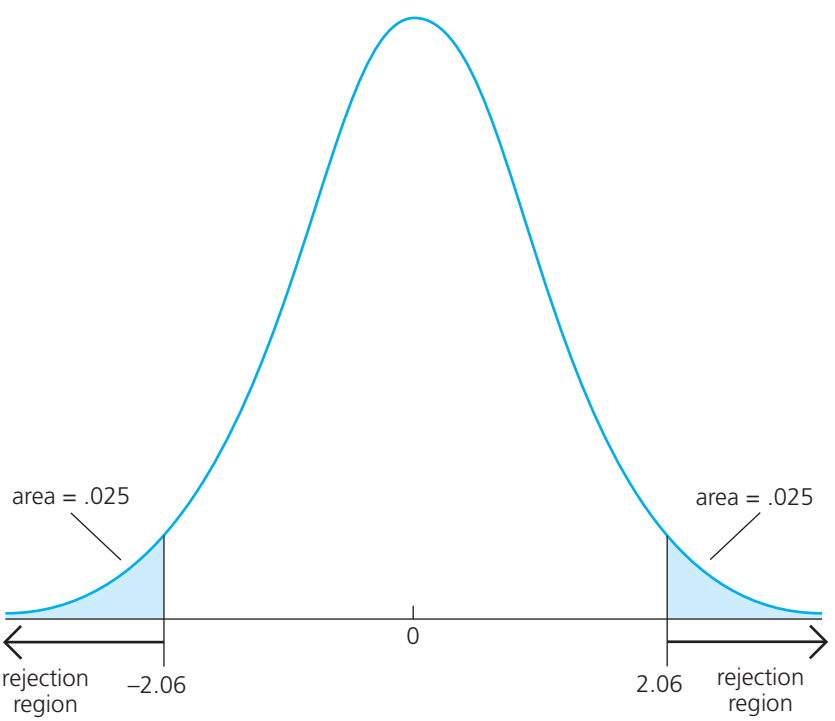
4-2b Two-Sided Alternatives

In applications, it is common to test the null hypothesis H_0: $\beta_j = 0$ against a **two-sided alternative**; that is,

$$H_1: \beta_j \neq 0. \qquad [4.10]$$

Under this alternative, x_j has a ceteris paribus effect on y without specifying whether the effect is positive or negative. This is the relevant alternative when the sign of β_j is not well determined by theory (or common sense). Even when we know whether β_j is positive or negative under the alternative, a two-sided test is often prudent. At a minimum, using a two-sided alternative prevents us from looking at the estimated equation and then basing the alternative on whether $\hat{\beta}_j$ is positive or negative. Using the regression estimates to help us formulate the null or alternative hypotheses is not allowed

Figure 4.4 5% rejection rule for the alternative H_1: $\beta_j \neq 0$ with 25 *df*.



because classical statistical inference presumes that we state the null and alternative about the population before looking at the data. For example, we should not first estimate the equation relating math performance to enrollment, note that the estimated effect is negative, and then decide the relevant alternative is H_1: $\beta_{enroll} < 0$.

When the alternative is two-sided, we are interested in the *absolute value* of the t statistic. The rejection rule for H_0: $\beta_j = 0$ against (4.10) is

$$|t_{\hat{\beta}_j}| > c, \qquad [4.11]$$

where $|\cdot|$ denotes absolute value and c is an appropriately chosen critical value. To find c, we again specify a significance level, say 5%. For a **two-tailed test**, c is chosen to make the area in each tail of the t distribution an equal 2.5%. In other words, c is the 97.5th percentile in the t distribution with $n - k - 1$ degrees of freedom. When $n - k - 1 = 25$, the 5% critical value for a two-sided test is $c = 2.060$. Figure 4.4 provides an illustration of this distribution.

When a specific alternative is not stated, it is usually considered to be two-sided. In the remainder of this text, the default will be a two-sided alternative, and 5% will be the default significance level. When carrying out empirical econometric analysis, it is always a good idea to be explicit about the alternative and the significance level. If H_0 is rejected in favor of (4.10) at the 5% level, we usually say that "x_j is **statistically significant**, or statistically different from zero, at the 5% level." If H_0 is not rejected, we say that "x_j is **statistically insignificant** at the 5% level."

EXAMPLE 4.3 Determinants of College GPA

We use the data in GPA1 to estimate a model explaining college GPA (*colGPA*), with the average number of lectures missed per week (*skipped*) as an additional explanatory variable. The estimated model is

$$\widehat{colGPA} = 1.39 + .412\, hsGPA + .015\, ACT - .083\, skipped$$
$$\qquad\qquad (.33)\quad (.094)\qquad\quad (.011)\qquad (.026)$$
$$n = 141,\ R^2 = .234.$$

We can easily compute t statistics to see which variables are statistically significant, using a two-sided alternative in each case. The 5% critical value is about 1.96, because the degrees of freedom $(141 - 4 = 137)$ is large enough to use the standard normal approximation. The 1% critical value is about 2.58.

The t statistic on *hsGPA* is 4.38, which is significant at very small significance levels. Thus, we say that "*hsGPA* is statistically significant at any *conventional* significance level." The t statistic on *ACT* is 1.36, which is not statistically significant at the 10% level against a two-sided alternative. The coefficient on *ACT* is also practically small: a 10-point increase in *ACT*, which is large, is predicted to increase *colGPA* by only .15 points. Thus, the variable *ACT* is practically, as well as statistically, insignificant.

The coefficient on *skipped* has a t statistic of $-.083/.026 = -3.19$, so *skipped* is statistically significant at the 1% significance level $(3.19 > 2.58)$. This coefficient means that another lecture missed per week lowers predicted *colGPA* by about .083. Thus, holding *hsGPA* and *ACT* fixed, the predicted difference in *colGPA* between a student who misses no lectures per week and a student who misses five lectures per week is about .42. Remember that this says nothing about specific students; rather, .42 is the estimated average across a subpopulation of students.

In this example, for each variable in the model, we could argue that a one-sided alternative is appropriate. The variables *hsGPA* and *skipped* are very significant using a two-tailed test and have the signs that we expect, so there is no reason to do a one-tailed test. On the other hand, against a one-sided alternative $(\beta_3 > 0)$, *ACT* is significant at the 10% level but not at the 5% level. This does not change the fact that the coefficient on *ACT* is pretty small.

4-2c Testing Other Hypotheses about β_j

Although H_0: $\beta_j = 0$ is the most common hypothesis, we sometimes want to test whether β_j is equal to some other given constant. Two common examples are $\beta_j = 1$ and $\beta_j = -1$. Generally, if the null is stated as

$$H_0: \beta_j = a_j, \tag{4.12}$$

where a_j is our hypothesized value of β_j, then the appropriate t statistic is

$$t = (\hat{\beta}_j - a_j)/\text{se}(\hat{\beta}_j).$$

As before, t measures how many estimated standard deviations $\hat{\beta}_j$ is away from the hypothesized value of β_j. The general t statistic is usefully written as

$$t = \frac{(\textit{estimate} - \textit{hypothesized value})}{\textit{standard error}}. \tag{4.13}$$

Under (4.12), this t statistic is distributed as t_{n-k-1} from Theorem 4.2. The usual t statistic is obtained when $a_j = 0$.

We can use the general t statistic to test against one-sided or two-sided alternatives. For example, if the null and alternative hypotheses are H_0: $\beta_j = 1$ and H_1: $\beta_j > 1$, then we find the critical value for a one-sided alternative *exactly* as before: the difference is in how we compute the t statistic, not in how we obtain the appropriate c. We reject H_0 in favor of H_1 if $t > c$. In this case, we would say that "$\hat{\beta}_j$ is statistically greater than one" at the appropriate significance level.

EXAMPLE 4.4 Campus Crime and Enrollment

Consider a simple model relating the annual number of crimes on college campuses (*crime*) to student enrollment (*enroll*):

$$\log(\textit{crime}) = \beta_0 + \beta_1 \log(\textit{enroll}) + u.$$

This is a constant elasticity model, where β_1 is the elasticity of crime with respect to enrollment. It is not much use to test H_0: $\beta_1 = 0$, as we expect the total number of crimes to increase as the size of the campus increases. A more interesting hypothesis to test would be that the elasticity of crime with respect to enrollment is one: H_0: $\beta_1 = 1$. This means that a 1% increase in enrollment leads to, on average, a 1% increase in crime. A noteworthy alternative is H_1: $\beta_1 > 1$, which implies that a 1% increase in enrollment increases campus crime by *more* than 1%. If $\beta_1 > 1$, then, in a relative sense—not just an absolute sense—crime is more of a problem on larger campuses. One way to see this is to take the exponential of the equation:

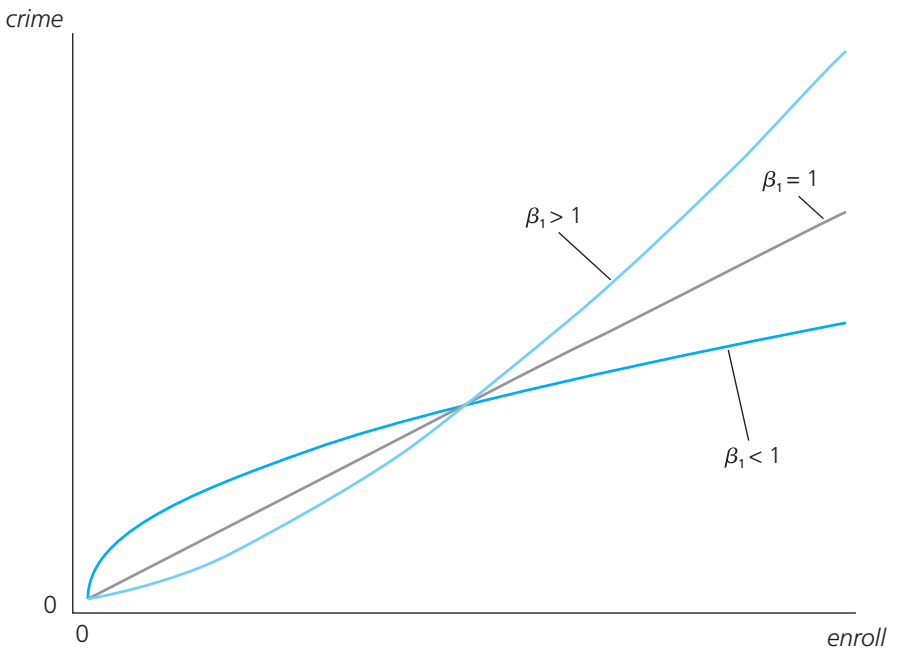
$$\textit{crime} = \exp(\beta_0)\textit{enroll}^{\beta_1}\exp(u).$$

(See Math Refresher A for properties of the natural logarithm and exponential functions.) For $\beta_0 = 0$ and $u = 0$, this equation is graphed in Figure 4.5 for $\beta_1 < 1$, $\beta_1 = 1$, and $\beta_1 > 1$.

We test $\beta_1 = 1$ against $\beta_1 > 1$ using data on 97 colleges and universities in the United States for the year 1992, contained in the data file CAMPUS. The data come from the FBI's *Uniform Crime Reports*, and the average number of campus crimes in the sample is about 394, while the average enrollment is about 16,076. The estimated equation (with estimates and standard errors rounded to two decimal places) is

$$\widehat{\log(\textit{crime})} = -6.63 + 1.27 \log(\textit{enroll}) \underset{(1.03)\ \ (0.11)}{}$$
$$n = 97, R^2 = .585. \tag{4.14}$$

Figure 4.5 Graph of *crime* = *enroll*$^{\beta_1}$ for $\beta_1 < 1$, $\beta_1 = 1$, and $\beta_1 > 1$.



The estimated elasticity of *crime* with respect to *enroll*, 1.27, is in the direction of the alternative $\beta_1 > 1$. But is there enough evidence to conclude that $\beta_1 > 1$? We need to be careful in testing this hypothesis, especially because the statistical output of standard regression packages is much more complex than the simplified output reported in equation (4.14). Our first instinct might be to construct "the" t statistic by taking the coefficient on log(*enroll*) and dividing it by its standard error, which is the t statistic reported by a regression package. But this is the *wrong* statistic for testing H_0: $\beta_1 = 1$. The correct t statistic is obtained from (4.13): we subtract the hypothesized value, unity, from the estimate and divide the result by the standard error of $\hat{\beta}_1$: $t = (1.27 - 1)/.11 = .27/.11 \approx 2.45$. The one-sided 5% critical value for a t distribution with $97 - 2 = 95$ df is about 1.66 (using $df = 120$), so we clearly reject $\beta_1 = 1$ in favor of $\beta_1 > 1$ at the 5% level. In fact, the 1% critical value is about 2.37, and so we reject the null in favor of the alternative at even the 1% level.

We should keep in mind that this analysis holds no other factors constant, so the elasticity of 1.27 is not necessarily a good estimate of ceteris paribus effect. It could be that larger enrollments are correlated with other factors that cause higher crime: larger schools might be located in higher crime areas. We could control for this by collecting data on crime rates in the local city.

For a two-sided alternative, for example H_0: $\beta_j = -1$, H_1: $\beta_1 \neq -1$, we still compute the t statistic as in (4.13): $t = (\hat{\beta}_j + 1)/\text{se}(\hat{\beta}_j)$ (notice how subtracting -1 means adding 1). The rejection rule is the usual one for a two-sided test: reject H_0 if $|t| > c$, where c is a two-tailed critical value. If H_0 is rejected, we say that "$\hat{\beta}_j$ is statistically different from negative one" at the appropriate significance level.

EXAMPLE 4.5 **Housing Prices and Air Pollution**

For a sample of 506 communities in the Boston area, we estimate a model relating median housing price (*price*) in the community to various community characteristics: *nox* is the amount of nitrogen oxide in the air, in parts per million; *dist* is a weighted distance of the community from five employment centers, in miles; *rooms* is the average number of rooms in houses in the community; and *stratio* is the average student-teacher ratio of schools in the community. The population model is

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 stratio + u.$$

Thus, β_1 is the elasticity of *price* with respect to *nox*. We wish to test H_0: $\beta_1 = -1$ against the alternative H_1: $\beta_1 \neq -1$. The t statistic for doing this test is $t = (\hat{\beta}_1 + 1)/\text{se}(\hat{\beta}_1)$.

Using the data in HPRICE2, the estimated model is

$$\widehat{\log(price)} = \underset{(0.32)}{11.08} - \underset{(.117)}{.954}\log(nox) - \underset{(.043)}{.134}\log(dist) + \underset{(.019)}{.255}rooms - \underset{(.006)}{.052}stratio$$
$$n = 506, R^2 = .581.$$

The slope estimates all have the anticipated signs. Each coefficient is statistically different from zero at very small significance levels, including the coefficient on log(*nox*). But we do not want to test that $\beta_1 = 0$. The null hypothesis of interest is H_0: $\beta_1 = -1$, with corresponding t statistic $(-.954 + 1)/.117 = .393$. There is little need to look in the t table for a critical value when the t statistic is this small: the estimated elasticity is not statistically different from -1 even at very large significance levels. Controlling for the factors we have included, there is little evidence that the elasticity is different from -1.

4-2d Computing *p*-Values for *t* Tests

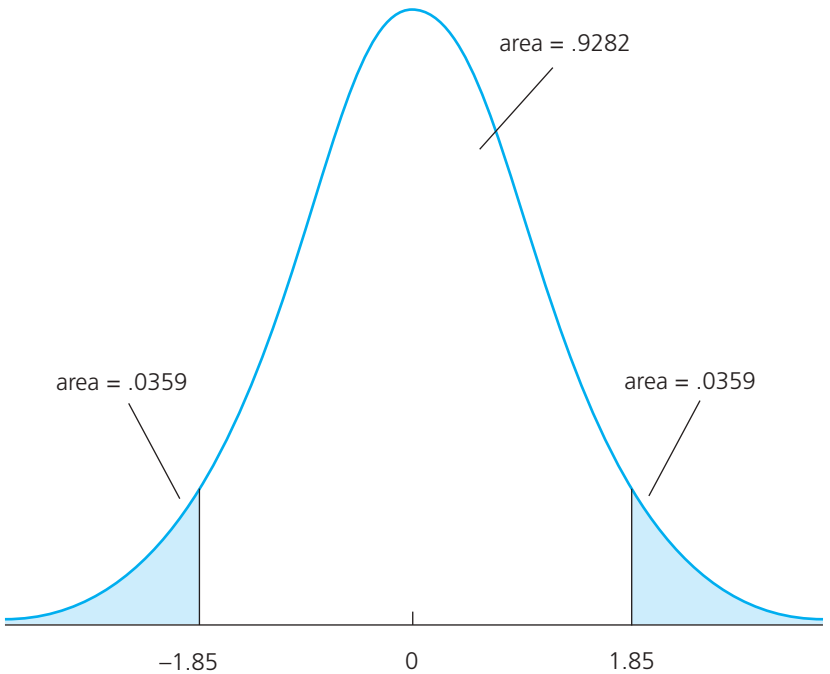
So far, we have talked about how to test hypotheses using a classical approach: after stating the alternative hypothesis, we choose a significance level, which then determines a critical value. Once the critical value has been identified, the value of the t statistic is compared with the critical value, and the null is either rejected or not rejected at the given significance level.

Even after deciding on the appropriate alternative, there is a component of arbitrariness to the classical approach, which results from having to choose a significance level ahead of time. Different researchers prefer different significance levels, depending on the particular application. There is no "correct" significance level.

Committing to a significance level ahead of time can hide useful information about the outcome of a hypothesis test. For example, suppose that we wish to test the null hypothesis that a parameter is zero against a two-sided alternative, and with 40 degrees of freedom we obtain a t statistic equal to 1.85. The null hypothesis is not rejected at the 5% level, because the t statistic is less than the two-tailed critical value of $c = 2.021$. A researcher whose agenda is not to reject the null could simply report this outcome along with the estimate: the null hypothesis is not rejected at the 5% level. Of course, if the t statistic, or the coefficient and its standard error, are reported, then we can also determine that the null hypothesis would be rejected at the 10% level, because the 10% critical value is $c = 1.684$.

Rather than testing at different significance levels, it is more informative to answer the following question: given the observed value of the t statistic, what is the *smallest* significance level at which the null hypothesis would be rejected? This level is known as the ***p*-value** for the test (see Math Refresher C). In the previous example, we know the p-value is greater than .05, because the null is not rejected at the 5% level, and we know that the p-value is less than .10, because the null is rejected at the 10% level. We obtain the actual p-value by computing the probability that a t random variable, with 40 *df*, is larger than 1.85 in absolute value. That is, the p-value is the significance level of the test when we use the value of the test statistic, 1.85 in the above example, as the critical value for the test. This p-value is shown in Figure 4.6.

Figure 4.6 Obtaining the *p*-value against a two-sided alternative, when $t = 1.85$ and $df = 40$.



Because a *p*-value is a probability, its value is always between zero and one. In order to compute *p*-values, we either need extremely detailed printed tables of the *t* distribution—which is not very practical—or a computer program that computes areas under the probability density function of the *t* distribution. Most modern regression packages have this capability. Some packages compute *p*-values routinely with each OLS regression, but only for certain hypotheses. If a regression package reports a *p*-value along with the standard OLS output, it is almost certainly the *p*-value for testing the null hypothesis H_0: $\beta_j = 0$ against the two-sided alternative. The *p*-value in this case is

$$P(|T| > |t|), \qquad [4.15]$$

where, for clarity, we let T denote a *t* distributed random variable with $n - k - 1$ degrees of freedom and let t denote the numerical value of the test statistic.

The *p*-value nicely summarizes the strength or weakness of the empirical evidence against the null hypothesis. Perhaps its most useful interpretation is the following: the *p*-value is the probability of observing a *t* statistic as extreme as we did *if the null hypothesis is true*. This means that *small* *p*-values are evidence *against* the null; large *p*-values provide little evidence against H_0. For example, if the *p*-value $= .50$ (reported always as a decimal, not a percentage), then we would observe a value of the *t* statistic as extreme as we did in 50% of all random samples when the null hypothesis is true; this is pretty weak evidence against H_0.

In the example with $df = 40$ and $t = 1.85$, the *p*-value is computed as

$$p\text{-value} = P(|T| > 1.85) = 2P(T > 1.85) = 2(.0359) = .0718,$$

where $P(T > 1.85)$ is the area to the right of 1.85 in a *t* distribution with 40 *df*. (This value was computed using the econometrics package Stata; it is not available in Table G.2.) This means that, if the

null hypothesis is true, we would observe an absolute value of the t statistic as large as 1.85 about 7.2 percent of the time. This provides some evidence against the null hypothesis, but we would not reject the null at the 5% significance level.

The previous example illustrates that once the p-value has been computed, a classical test can be carried out at any desired level. If α denotes the significance level of the test (in decimal form), then H_0 is rejected if p-value $< \alpha$; otherwise, H_0 is not rejected at the $100\cdot\alpha\%$ level.

Computing p-values for one-sided alternatives is also quite simple. Suppose, for example, that we test H_0: $\beta_j = 0$ against H_1: $\beta_j > 0$. If $\hat{\beta}_j < 0$, then computing a p-value is not important: we know that the p-value is greater than .50, which will never cause us to reject H_0 in favor of H_1. If $\hat{\beta}_j > 0$, then $t > 0$ and the p-value is just the probability that a random t variable with the appropriate df exceeds the value t. Some regression packages only compute p-values for two-sided alternatives. But it is simple to obtain the one-sided p-value: just divide the two-sided p-value by 2.

If the alternative is H_1: $\beta_j < 0$, it makes sense to compute a p-value if $\hat{\beta}_j < 0$ (and hence $t < 0$): p-value $= P(T < t) = P(T > |t|)$ because the t distribution is symmetric about zero. Again, this can be obtained as one-half of the p-value for the two-tailed test.

GOING FURTHER 4.3

Suppose you estimate a regression model and obtain $\hat{\beta}_1 = .56$ and p-value $= .086$ for testing H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$. What is the p-value for testing H_0: $\beta_1 = 0$ against H_1: $\beta_1 > 0$?

Because you will quickly become familiar with the magnitudes of t statistics that lead to statistical significance, especially for large sample sizes, it is not always crucial to report p-values for t statistics. But it does not hurt to report them. Further, when we discuss F testing in Section 4-5, we will see that it is important to compute p-values, because critical values for F tests are not so easily memorized.

4-2e A Reminder on the Language of Classical Hypothesis Testing

When H_0 is not rejected, we prefer to use the language "we fail to reject H_0 at the x% level," rather than "H_0 is accepted at the x% level." We can use Example 4.5 to illustrate why the former statement is preferred. In this example, the estimated elasticity of *price* with respect to *nox* is $-.954$, and the t statistic for testing H_0: $\beta_{nox} = -1$ is $t = .393$; therefore, we cannot reject H_0. But there are many other values for β_{nox} (more than we can count) that cannot be rejected. For example, the t statistic for H_0: $\beta_{nox} = -.9$ is $(-.954 + .9)/.117 = -.462$, and so this null is not rejected either. Clearly $\beta_{nox} = -1$ and $\beta_{nox} = -.9$ cannot both be true, so it makes no sense to say that we "accept" either of these hypotheses. All we can say is that the data do not allow us to reject either of these hypotheses at the 5% significance level.

4-2f Economic, or Practical, versus Statistical Significance

Because we have emphasized *statistical significance* throughout this section, now is a good time to remember that we should pay attention to the magnitude of the *coefficient* estimates in addition to the size of the t statistics. The statistical significance of a variable x_j is determined entirely by the size of $t_{\hat{\beta}_j}$, whereas the **economic significance** or **practical significance** of a variable is related to the size (and sign) of $\hat{\beta}_j$.

Recall that the t statistic for testing H_0: $\beta_j = 0$ is defined by dividing the estimate by its standard error: $t_{\hat{\beta}_j} = \hat{\beta}_j/\text{se}(\hat{\beta}_j)$. Thus, $t_{\hat{\beta}_j}$ can indicate statistical significance either because $\hat{\beta}_j$ is "large" or because $\text{se}(\hat{\beta}_j)$ is "small." It is important in practice to distinguish between these reasons for statistically significant t statistics. Too much focus on statistical significance can lead to the false conclusion that a variable is "important" for explaining y even though its estimated effect is modest.

EXAMPLE 4.6 **Participation Rates in 401(k) Plans**

In Example 3.3, we used the data on 401(k) plans to estimate a model describing participation rates in terms of the firm's match rate and the age of the plan. We now include a measure of firm size, the total number of firm employees (*totemp*). The estimated equation is

$$\widehat{prate} = 80.29 + 5.44\,mrate + .269\,age - .00013\,totemp$$
$$(0.78)\ (0.52)\qquad(.045)\qquad(.00004)$$
$$n = 1{,}534,\ R^2 = .100.$$

The smallest t statistic in absolute value is that on the variable *totemp*: $t = -.00013/.00004 = -3.25$, and this is statistically significant at very small significance levels. (The two-tailed p-value for this t statistic is about .001.) Thus, all of the variables are statistically significant at rather small significance levels.

How big, in a practical sense, is the coefficient on *totemp*? Holding *mrate* and *age* fixed, if a firm grows by 10,000 employees, the participation rate falls by $10{,}000(.00013) = 1.3$ percentage points. This is a huge increase in number of employees with only a modest effect on the participation rate. Thus, although firm size does affect the participation rate, the effect is not practically very large.

The previous example shows that it is especially important to interpret the magnitude of the coefficient, in addition to looking at t statistics, when working with large samples. With large sample sizes, parameters can be estimated very precisely: standard errors are often quite small relative to the coefficient estimates, which usually results in statistical significance.

Some researchers insist on using smaller significance levels as the sample size increases, partly as a way to offset the fact that standard errors are getting smaller. For example, if we feel comfortable with a 5% level when n is a few hundred, we might use the 1% level when n is a few thousand. Using a smaller significance level means that economic and statistical significance are more likely to coincide, but there are no guarantees: in the previous example, even if we use a significance level as small as .1% (one-tenth of 1%), we would still conclude that *totemp* is statistically significant.

Many researchers are also willing to entertain larger significance levels in applications with small sample sizes, reflecting the fact that it is harder to find significance with smaller sample sizes. (Smaller sample sizes lead to less precise estimators, and the critical values are larger in magnitude, two factors that make it harder to find statistical significance.) Unfortunately, one's willingness to consider higher significance levels can depend on one's underlying agenda.

EXAMPLE 4.7 **Effect of Job Training on Firm Scrap Rates**

The scrap rate for a manufacturing firm is the number of defective items—products that must be discarded—out of every 100 produced. Thus, for a given number of items produced, a decrease in the scrap rate reflects higher worker productivity.

We can use the scrap rate to measure the effect of worker training on productivity. Using the data in JTRAIN, but only for the year 1987 and for nonunionized firms, we obtain the following estimated equation:

$$\widehat{\log(scrap)} = 12.46 - .029\,hrsemp - .962\log(sales) + .761\log(employ)$$
$$(5.69)\ (.023)\qquad(.453)\qquad(.407)$$
$$n = 29,\ R^2 = .262.$$

The variable *hrsemp* is annual hours of training per employee, *sales* is annual firm sales (in dollars), and *employ* is the number of firm employees. For 1987, the average scrap rate in the sample is about 4.6 and the average of *hrsemp* is about 8.9.

The main variable of interest is *hrsemp*. One more hour of training per employee lowers log(*scrap*) by .029, which means the scrap rate is about 2.9% lower. Thus, if *hrsemp* increases by 5—each employee is trained 5 more hours per year—the scrap rate is estimated to fall by $5(2.9) = 14.5\%$. This seems like a reasonably large effect, but whether the additional training is worthwhile to the firm depends on the cost of training and the benefits from a lower scrap rate. We do not have the numbers needed to do a cost-benefit analysis, but the estimated effect seems nontrivial.

What about the *statistical significance* of the training variable? The t statistic on *hrsemp* is $-.029/.023 = -1.26$, and now you probably recognize this as not being large enough in magnitude to conclude that *hrsemp* is statistically significant at the 5% level. In fact, with $29 - 4 = 25$ degrees of freedom for the one-sided alternative, H_1: $\beta_{hrsemp} < 0$, the 5% critical value is about -1.71. Thus, using a strict 5% level test, we must conclude that *hrsemp* is not statistically significant, even using a one-sided alternative.

Because the sample size is pretty small, we might be more liberal with the significance level. The 10% critical value is -1.32, and so *hrsemp* is almost significant against the one-sided alternative at the 10% level. The p-value is easily computed as $P(T_{25} < -1.26) = .110$. This may be a low enough p-value to conclude that the estimated effect of training is not just due to sampling error, but opinions would legitimately differ on whether a one-sided p-value of .11 is sufficiently small.

Remember that large standard errors can also be a result of multicollinearity (high correlation among some of the independent variables), even if the sample size seems fairly large. As we discussed in Section 3-4, there is not much we can do about this problem other than to collect more data or change the scope of the analysis by dropping or combining certain independent variables. As in the case of a small sample size, it can be hard to precisely estimate partial effects when some of the explanatory variables are highly correlated. (Section 4-5 contains an example.)

We end this section with some guidelines for discussing the economic and statistical significance of a variable in a multiple regression model:

1. Check for statistical significance. If the variable is statistically significant, discuss the magnitude of the coefficient to get an idea of its practical or economic importance. This latter step can require some care, depending on how the independent and dependent variables appear in the equation. (In particular, what are the units of measurement? Do the variables appear in logarithmic form?)
2. If a variable is not statistically significant at the usual levels (10%, 5%, or 1%), you might still ask if the variable has the expected effect on y and whether that effect is practically large. If it is large, you should compute a p-value for the t statistic. For small sample sizes, you can sometimes make a case for p-values as large as .20 (but there are no hard rules). With large p-values, that is, small t statistics, we are treading on thin ice because the practically large estimates may be due to sampling error: a different random sample could result in a very different estimate.
3. It is common to find variables with small t statistics that have the "wrong" sign. For practical purposes, these can be ignored: we conclude that the variables are statistically insignificant. A significant variable that has the unexpected sign and a practically large effect is much more troubling and difficult to resolve. One must usually think more about the model and the nature of the data to solve such problems. Often, a counterintuitive, significant estimate results from the omission of a key variable or from one of the important problems we will discuss in Chapters 9 and 15.

4-3 Confidence Intervals

Under the CLM assumptions, we can easily construct a **confidence interval (CI)** for the population parameter β_j. Confidence intervals are also called *interval estimates* because they provide a range of likely values for the population parameter, and not just a point estimate.

Using the fact that $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ has a t distribution with $n - k - 1$ degrees of freedom [see (4.3)], simple manipulation leads to a CI for the unknown β_j: a *95% confidence interval*, given by

$$\hat{\beta}_j \pm c \cdot \text{se}(\hat{\beta}_j), \qquad [4.16]$$

where the constant c is the 97.5th percentile in a t_{n-k-1} distribution. More precisely, the lower and upper bounds of the confidence interval are given by

$$\underline{\beta}_j \equiv \hat{\beta}_j - c \cdot \text{se}(\hat{\beta}_j)$$

and

$$\overline{\beta}_j \equiv \hat{\beta}_j + c \cdot \text{se}(\hat{\beta}_j),$$

respectively.

At this point, it is useful to review the meaning of a confidence interval. If random samples were obtained over and over again, with $\underline{\beta}_j$ and $\overline{\beta}_j$ computed each time, then the (unknown) population value β_j would lie in the interval $(\underline{\beta}_j, \overline{\beta}_j)$ for 95% of the samples. Unfortunately, for the single sample that we use to construct the CI, we do not know whether β_j is actually contained in the interval. We hope we have obtained a sample that is one of the 95% of all samples where the interval estimate contains β_j, but we have no guarantee.

Constructing a confidence interval is very simple when using current computing technology. Three quantities are needed: $\hat{\beta}_j$, $\text{se}(\hat{\beta}_j)$, and c. The coefficient estimate and its standard error are reported by any regression package. To obtain the value c, we must know the degrees of freedom, $n - k - 1$, and the level of confidence—95% in this case. Then, the value for c is obtained from the t_{n-k-1} distribution.

As an example, for $df = n - k - 1 = 25$, a 95% confidence interval for any β_j is given by $[\hat{\beta}_j - 2.06 \cdot \text{se}(\hat{\beta}_j), \hat{\beta}_j + 2.06 \cdot \text{se}(\hat{\beta}_j)]$.

When $n - k - 1 > 120$, the t_{n-k-1} distribution is close enough to normal to use the 97.5th percentile in a standard normal distribution for constructing a 95% CI: $\hat{\beta}_j \pm 1.96 \cdot \text{se}(\hat{\beta}_j)$. In fact, when $n - k - 1 > 50$, the value of c is so close to 2 that we can use a simple *rule of thumb* for a 95% confidence interval: $\hat{\beta}_j$ plus or minus two of its standard errors. For small degrees of freedom, the exact percentiles should be obtained from the t tables.

It is easy to construct confidence intervals for any other level of confidence. For example, a 90% CI is obtained by choosing c to be the 95th percentile in the t_{n-k-1} distribution. When $df = n - k - 1 = 25$, $c = 1.71$, and so the 90% CI is $\hat{\beta}_j \pm 1.71 \cdot \text{se}(\hat{\beta}_j)$, which is necessarily narrower than the 95% CI. For a 99% CI, c is the 99.5th percentile in the t_{25} distribution. When $df = 25$, the 99% CI is roughly $\hat{\beta}_j \pm 2.79 \cdot \text{se}(\hat{\beta}_j)$, which is inevitably wider than the 95% CI.

Many modern regression packages save us from doing any calculations by reporting a 95% CI along with each coefficient and its standard error. After a confidence interval is constructed, it is easy to carry out two-tailed hypotheses tests. If the null hypothesis is H_0: $\beta_j = a_j$, then H_0 is rejected against H_1: $\beta_j \neq a_j$ at (say) the 5% significance level if, and only if, a_j is *not* in the 95% confidence interval.

EXAMPLE 4.8 Model of R&D Expenditures

Economists studying industrial organization are interested in the relationship between firm size—often measured by annual sales—and spending on research and development (R&D). Typically, a constant elasticity model is used. One might also be interested in the ceteris paribus effect of the profit margin—that is, profits as a percentage of sales—on R&D spending. Using the data in RDCHEM on 32 U.S. firms in the chemical industry, we estimate the following equation (with standard errors in parentheses below the coefficients):

$$\widehat{\log(rd)} = \underset{(.47)}{-4.38} + \underset{(.060)}{1.084}\log(sales) + \underset{(.0128)}{.0217}\,profmarg$$
$$n = 32,\ R^2 = .918.$$

The estimated elasticity of R&D spending with respect to firm sales is 1.084, so that, holding profit margin fixed, a 1% increase in sales is associated with a 1.084% increase in R&D spending. (Incidentally, R&D and sales are both measured in millions of dollars, but their units of measurement have no effect on the elasticity estimate.) We can construct a 95% confidence interval for the sales elasticity once we note that the estimated model has $n - k - 1 = 32 - 2 - 1 = 29$ degrees of freedom. From Table G.2, we find the 97.5th percentile in a t_{29} distribution: $c = 2.045$. Thus, the 95% confidence interval for $\beta_{\log(sales)}$ is $1.084 \pm .060(2.045)$, or about (.961,1.21). That zero is well outside this interval is hardly surprising: we expect R&D spending to increase with firm size. More interesting is that unity is included in the 95% confidence interval for $\beta_{\log(sales)}$, which means that we cannot reject H_0: $\beta_{\log(sales)} = 1$ against H_1: $\beta_{\log(sales)} \neq 1$ at the 5% significance level. In other words, the estimated R&D-sales elasticity is not statistically different from 1 at the 5% level. (The estimate is not practically different from 1, either.)

The estimated coefficient on *profmarg* is also positive, and the 95% confidence interval for the population parameter, $\beta_{profmarg}$, is $.0217 \pm .0128(2.045)$, or about $(-.0045, .0479)$. In this case, zero is included in the 95% confidence interval, so we fail to reject H_0: $\beta_{profmarg} = 0$ against H_1: $\beta_{profmarg} \neq 0$ at the 5% level. Nevertheless, the t statistic is about 1.70, which gives a two-sided p-value of about .10, and so we would conclude that *profmarg* is statistically significant at the 10% level against the two-sided alternative, or at the 5% level against the one-sided alternative H_1: $\beta_{profmarg} > 0$. Plus, the economic size of the profit margin coefficient is not trivial: holding *sales* fixed, a one percentage point increase in *profmarg* is estimated to increase R&D spending by $100(.0217) \approx 2.2\%$. A complete analysis of this example goes beyond simply stating whether a particular value, zero in this case, is or is not in the 95% confidence interval.

You should remember that a confidence interval is only as good as the underlying assumptions used to construct it. If we have omitted important factors that are correlated with the explanatory variables, then the coefficient estimates are not reliable: OLS is biased. If heteroskedasticity is present—for instance, in the previous example, if the variance of $\log(rd)$ depends on any of the explanatory variables—then the standard error is not valid as an estimate of $\text{sd}(\hat{\beta}_j)$ (as we discussed in Section 3-4), and the confidence interval computed using these standard errors will not truly be a 95% CI. We have also used the normality assumption on the errors in obtaining these CIs, but, as we will see in Chapter 5, this is not as important for applications involving hundreds of observations.

4-4 Testing Hypotheses about a Single Linear Combination of the Parameters

The previous two sections have shown how to use classical hypothesis testing or confidence intervals to test hypotheses about a single β_j at a time. In applications, we must often test hypotheses involving more than one of the population parameters. In this section, we show how to test a single hypothesis involving more than one of the β_j. Section 4-5 shows how to test multiple hypotheses.

To illustrate the general approach, we will consider a simple model to compare the returns to education at junior colleges and four-year colleges; for simplicity, we refer to the latter as "universities." [Kane and Rouse (1995) provide a detailed analysis of the returns to two- and four-year colleges.] The population includes working people with a high school degree, and the model is

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u, \qquad [4.17]$$

where

jc = number of years attending a two-year college.
$univ$ = number of years at a four-year college.
$exper$ = months in the workforce.

Note that any combination of junior college and four-year college is allowed, including $jc = 0$ and $univ = 0$.

The hypothesis of interest is whether one year at a junior college is worth one year at a university: this is stated as

$$H_0: \beta_1 = \beta_2. \tag{4.18}$$

Under H_0, another year at a junior college and another year at a university lead to the same ceteris paribus percentage increase in *wage*. For the most part, the alternative of interest is one-sided: a year at a junior college is worth less than a year at a university. This is stated as

$$H_1: \beta_1 < \beta_2. \tag{4.19}$$

The hypotheses in (4.18) and (4.19) concern *two* parameters, β_1 and β_2, a situation we have not faced yet. We cannot simply use the individual t statistics for $\hat{\beta}_1$ and $\hat{\beta}_2$ to test H_0. However, conceptually, there is no difficulty in constructing a t statistic for testing (4.18). To do so, we rewrite the null and alternative as $H_0: \beta_1 - \beta_2 = 0$ and $H_1: \beta_1 - \beta_2 < 0$, respectively. The t statistic is based on whether the estimated difference $\hat{\beta}_1 - \hat{\beta}_2$ is sufficiently less than zero to warrant rejecting (4.18) in favor of (4.19). To account for the sampling error in our estimators, we standardize this difference by dividing by the standard error:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{se}(\hat{\beta}_1 - \hat{\beta}_2)}. \tag{4.20}$$

Once we have the t statistic in (4.20), testing proceeds as before. We choose a significance level for the test and, based on the *df*, obtain a critical value. Because the alternative is of the form in (4.19), the rejection rule is of the form $t < -c$, where c is a positive value chosen from the appropriate t distribution. Or we compute the t statistic and then compute the p-value (see Section 4-2).

The only thing that makes testing the equality of two different parameters more difficult than testing about a single β_j is obtaining the standard error in the denominator of (4.20). Obtaining the numerator is trivial once we have performed the OLS regression. Using the data in TWOYEAR, which comes from Kane and Rouse (1995), we estimate equation (4.17):

$$\begin{aligned} \widehat{\log(wage)} &= 1.472 + .0667\, jc + .0769\, univ + .0049\, exper \\ &\qquad (.021)\ \ (.0068)\qquad (.0023)\qquad\ (.0002) \\ n &= 6{,}763,\ R^2 = .222. \end{aligned} \tag{4.21}$$

It is clear from (4.21) that *jc* and *univ* have both economically and statistically significant effects on wage. This is certainly of interest, but we are more concerned about testing whether the estimated *difference* in the coefficients is statistically significant. The difference is estimated as $\hat{\beta}_1 - \hat{\beta}_2 = -.0102$, so the return to a year at a junior college is about one percentage point less than a year at a university. Economically, this is not a trivial difference. The difference of $-.0102$ is the numerator of the t statistic in (4.20).

Unfortunately, the regression results in equation (4.21) do *not* contain enough information to obtain the standard error of $\hat{\beta}_1 - \hat{\beta}_2$. It might be tempting to claim that $\text{se}(\hat{\beta}_1 - \hat{\beta}_2) = \text{se}(\hat{\beta}_1) - \text{se}(\hat{\beta}_2)$, but this is not true. In fact, if we reversed the roles of $\hat{\beta}_1$ and $\hat{\beta}_2$, we would wind up with a negative standard error of the difference using the difference in standard errors. Standard errors must *always* be positive because they are estimates of standard deviations. Although the standard error of the difference $\hat{\beta}_1 - \hat{\beta}_2$ certainly depends on $\text{se}(\hat{\beta}_1)$ and $\text{se}(\hat{\beta}_2)$, it does so in a somewhat complicated way. To find $\text{se}(\hat{\beta}_1 - \hat{\beta}_2)$, we first obtain the variance of the difference. Using the results on variances in Math Refresher B, we have

$$\text{Var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\,\text{Cov}(\hat{\beta}_1, \hat{\beta}_2). \tag{4.22}$$

Observe carefully how the two variances are *added* together, and twice the covariance is then subtracted. The standard deviation of $\hat{\beta}_1 - \hat{\beta}_2$ is just the square root of (4.22), and, because $[\text{se}(\hat{\beta}_1)]^2$ is an unbiased estimator of $\text{Var}(\hat{\beta}_1)$, and similarly for $[\text{se}(\hat{\beta}_2)]^2$, we have

$$\text{se}(\hat{\beta}_1 - \hat{\beta}_2) = \{[\text{se}(\hat{\beta}_1)]^2 + [\text{se}(\hat{\beta}_2)]^2 - 2s_{12}\}^{1/2}, \quad [4.23]$$

where s_{12} denotes an estimate of $\text{Cov}(\hat{\beta}_1,\hat{\beta}_2)$. We have not displayed a formula for $\text{Cov}(\hat{\beta}_1,\hat{\beta}_2)$. Some regression packages have features that allow one to obtain s_{12}, in which case one can compute the standard error in (4.23) and then the t statistic in (4.20). Advanced Treatment E shows how to use matrix algebra to obtain s_{12}.

Some of the more sophisticated econometrics programs include special commands that can be used for testing hypotheses about linear combinations. Here, we cover an approach that is simple to compute in virtually any statistical package. Rather than trying to compute $\text{se}(\hat{\beta}_1 - \hat{\beta}_2)$ from (4.23), it is much easier to estimate a different model that directly delivers the standard error of interest. Define a new parameter as the difference between β_1 and β_2: $\theta_1 = \beta_1 - \beta_2$. Then, we want to test

$$H_0: \theta_1 = 0 \text{ against } H_1: \theta_1 < 0. \quad [4.24]$$

The t statistic in (4.20) in terms of $\hat{\theta}_1$ is just $t = \hat{\theta}_1/\text{se}(\hat{\theta}_1)$. The challenge is finding $\text{se}(\hat{\theta}_1)$.

We can do this by rewriting the model so that θ_1 appears directly on one of the independent variables. Because $\theta_1 = \beta_1 - \beta_2$, we can also write $\beta_1 = \theta_1 + \beta_2$. Plugging this into (4.17) and rearranging gives the equation

$$\begin{aligned} \log(wage) &= \beta_0 + (\theta_1 + \beta_2)jc + \beta_2 univ + \beta_3 exper + u \\ &= \beta_0 + \theta_1 jc + \beta_2(jc + univ) + \beta_3 exper + u. \end{aligned} \quad [4.25]$$

The key insight is that the parameter we are interested in testing hypotheses about, θ_1, now multiplies the variable *jc*. The intercept is still β_0, and *exper* still shows up as being multiplied by β_3. More importantly, there is a new variable multiplying β_2, namely $jc + univ$. Thus, if we want to directly estimate θ_1 and obtain the standard error of $\hat{\theta}_1$, then we must construct the new variable $jc + univ$ and include it in the regression model in place of *univ*. In this example, the new variable has a natural interpretation: it is *total* years of college, so define $totcoll = jc + univ$ and write (4.25) as

$$\log(wage) = \beta_0 + \theta_1 jc + \beta_2 totcoll + \beta_3 exper + u. \quad [4.26]$$

The parameter β_1 has disappeared from the model, while θ_1 appears explicitly. This model is really just a different way of writing the original model. The only reason we have defined this new model is that, when we estimate it, the coefficient on *jc* is $\hat{\theta}_1$, and, more importantly, $\text{se}(\hat{\theta}_1)$ is reported along with the estimate. The t statistic that we want is the one reported by any regression package on the variable *jc* (*not* the variable *totcoll*).

When we do this with the 6,763 observations used earlier, the result is

$$\begin{aligned} \widehat{\log(wage)} &= 1.472 - .0102\ jc + .0769\ totcoll + .0049\ exper \\ &\qquad (.021)\ \ (.0069)\qquad (.0023)\qquad\ \ (.0002) \\ n &= 6{,}763,\ R^2 = .222. \end{aligned} \quad [4.27]$$

The only number in this equation that we could not get from (4.21) is the standard error for the estimate $-.0102$, which is .0069. The t statistic for testing (4.18) is $-.0102/.0069 = -1.48$. Against the one-sided alternative (4.19), the p-value is about .070, so there is some, but not strong, evidence against (4.18).

The intercept and slope estimate on *exper*, along with their standard errors, are the same as in (4.21). This fact *must* be true, and it provides one way of checking whether the transformed equation has been properly estimated. The coefficient on the new variable, *totcoll*, is the same as the coefficient on *univ* in (4.21), and the standard error is also the same. We know that this must happen by comparing (4.17) and (4.25).

It is quite simple to compute a 95% confidence interval for $\theta_1 = \beta_1 - \beta_2$. Using the standard normal approximation, the CI is obtained as usual: $\hat{\theta}_1 \pm 1.96\ \text{se}(\hat{\theta}_1)$, which in this case leads to $-.0102 \pm .0135$.

The strategy of rewriting the model so that it contains the parameter of interest works in all cases and is easy to implement. (See Computer Exercises C1 and C3 for other examples.)

4-5 Testing Multiple Linear Restrictions: The *F* Test

The *t* statistic associated with any OLS coefficient can be used to test whether the corresponding unknown parameter in the population is equal to any given constant (which is usually, but not always, zero). We have just shown how to test hypotheses about a single linear combination of the β_j by rearranging the equation and running a regression using transformed variables. But so far, we have only covered hypotheses involving a *single* restriction. Frequently, we wish to test *multiple* hypotheses about the underlying parameters $\beta_0, \beta_1, \ldots, \beta_k$. We begin with the leading case of testing whether a set of independent variables has no partial effect on a dependent variable.

4-5a Testing Exclusion Restrictions

We already know how to test whether a particular variable has no partial effect on the dependent variable: use the *t* statistic. Now, we want to test whether a *group* of variables has no effect on the dependent variable. More precisely, the null hypothesis is that a set of variables has no effect on *y*, once another set of variables has been controlled.

As an illustration of why testing significance of a group of variables is useful, we consider the following model that explains major league baseball players' salaries:

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u, \quad [4.28]$$

where *salary* is the 1993 total salary, *years* is years in the league, *gamesyr* is average games played per year, *bavg* is career batting average (for example, *bavg* = 250), *hrunsyr* is home runs per year, and *rbisyr* is runs batted in per year. Suppose we want to test the null hypothesis that, once years in the league and games per year have been controlled for, the statistics measuring performance—*bavg*, *hrunsyr*, and *rbisyr*—have no effect on salary. Essentially, the null hypothesis states that productivity as measured by baseball statistics has no effect on salary.

In terms of the parameters of the model, the null hypothesis is stated as

$$H_0: \beta_3 = 0, \beta_4 = 0, \beta_5 = 0. \quad [4.29]$$

The null (4.29) constitutes three **exclusion restrictions**: if (4.29) is true, then *bavg*, *hrunsyr*, and *rbisyr* have no effect on log(*salary*) after *years* and *gamesyr* have been controlled for and therefore should be excluded from the model. This is an example of a set of **multiple restrictions** because we are putting more than one restriction on the parameters in (4.28); we will see more general examples of multiple restrictions later. A test of multiple restrictions is called a **multiple hypotheses test** or a **joint hypotheses test**.

What should be the alternative to (4.29)? If what we have in mind is that "performance statistics matter, even after controlling for years in the league and games per year," then the appropriate alternative is simply

$$H_1: H_0 \text{ is not true.} \quad [4.30]$$

The alternative (4.30) holds if at least one of β_3, β_4, or β_5 is different from zero. (Any or all could be different from zero.) The test we study here is constructed to detect any violation of H_0. It is also valid when the alternative is something like $H_1: \beta_3 > 0$, or $\beta_4 > 0$, or $\beta_5 > 0$, but it will not be the best

possible test under such alternatives. We do not have the space or statistical background necessary to cover tests that have more power under multiple one-sided alternatives.

How should we proceed in testing (4.29) against (4.30)? It is tempting to test (4.29) by using the *t* statistics on the variables *bavg*, *hrunsyr*, and *rbisyr* to determine whether each variable is *individually* significant. This option is not appropriate. A particular *t* statistic tests a hypothesis that puts no restrictions on the other parameters. Besides, we would have three outcomes to contend with—one for each *t* statistic. What would constitute rejection of (4.29) at, say, the 5% level? Should all three or only one of the three *t* statistics be required to be significant at the 5% level? These are hard questions, and fortunately we do not have to answer them. Furthermore, using separate *t* statistics to test a multiple hypothesis like (4.29) can be very misleading. We need a way to test the exclusion restrictions *jointly*.

To illustrate these issues, we estimate equation (4.28) using the data in MLB1. This gives

$$\begin{aligned}\widehat{\log(salary)} &= 11.19 + .0689\ years + .0126\ gamesyr \\ &\qquad (0.29)\ \ (.0121)\qquad\ (.0026) \\ &\quad + .00098\ bavg + .0144\ hrunsyr + .0108\ rbisyr \\ &\qquad (.00110)\qquad (.0161)\qquad\ (.0072) \\ n &= 353,\ \text{SSR} = 183.186,\ R^2 = .6278,\end{aligned} \tag{4.31}$$

where SSR is the sum of squared residuals. (We will use this later.) We have left several terms after the decimal in SSR and *R*-squared to facilitate future comparisons. Equation (4.31) reveals that, whereas *years* and *gamesyr* are statistically significant, none of the variables *bavg*, *hrunsyr*, and *rbisyr* has a statistically significant *t* statistic against a two-sided alternative, at the 5% significance level. (The *t* statistic on *rbisyr* is the closest to being significant; its two-sided *p*-value is .134.) Thus, based on the three *t* statistics, it appears that we cannot reject H_0.

This conclusion turns out to be wrong. To see this, we must derive a test of multiple restrictions whose distribution is known and tabulated. The sum of squared residuals now turns out to provide a very convenient basis for testing multiple hypotheses. We will also show how the *R*-squared can be used in the special case of testing for exclusion restrictions.

Knowing the sum of squared residuals in (4.31) tells us nothing about the truth of the hypothesis in (4.29). However, the factor that will tell us something is how much the SSR increases when we drop the variables *bavg*, *hrunsyr*, and *rbisyr* from the model. Remember that, because the OLS estimates are chosen to minimize the sum of squared residuals, the SSR *always* increases when variables are dropped from the model; this is an algebraic fact. The question is whether this increase is large enough, *relative* to the SSR in the model with all of the variables, to warrant rejecting the null hypothesis.

The model without the three variables in question is simply

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + u. \tag{4.32}$$

In the context of hypothesis testing, equation (4.32) is the **restricted model** for testing (4.29); model (4.28) is called the **unrestricted model**. The restricted model always has fewer parameters than the unrestricted model.

When we estimate the restricted model using the data in MLB1, we obtain

$$\begin{aligned}\widehat{\log(salary)} &= 11.22 + .0713\ years + .0202\ gamesyr \\ &\qquad (.11)\ \ (.0125)\qquad\ (.0013) \\ n &= 353,\ \text{SSR} = 198.311,\ R^2 = .5971.\end{aligned} \tag{4.33}$$

As we surmised, the SSR from (4.33) is greater than the SSR from (4.31), and the *R*-squared from the restricted model is less than the *R*-squared from the unrestricted model. What we need to decide is whether the increase in the SSR in going from the unrestricted model to the restricted model (183.186 to 198.311) is large enough to warrant rejection of (4.29). As with all testing, the answer depends on the significance level of the test. But we cannot carry out the test at a chosen significance level until we

have a statistic whose distribution is known, and can be tabulated, under H_0. Thus, we need a way to combine the information in the two SSRs to obtain a test statistic with a known distribution under H_0.

Because it is no more difficult, we might as well derive the test for the general case. Write the *unrestricted* model with k independent variables as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u; \quad [4.34]$$

the number of parameters in the unrestricted model is $k + 1$. (Remember to add one for the intercept.) Suppose that we have q exclusion restrictions to test: that is, the null hypothesis states that q of the variables in (4.34) have zero coefficients. For notational simplicity, assume that it is the last q variables in the list of independent variables: $x_{k-q+1}, \ldots, x_k$. (The order of the variables, of course, is arbitrary and unimportant.) The null hypothesis is stated as

$$H_0: \beta_{k-q+1} = 0, \ldots, \beta_k = 0, \quad [4.35]$$

which puts q exclusion restrictions on the model (4.34). The alternative to (4.35) is simply that it is false; this means that at least one of the parameters listed in (4.35) is different from zero. When we impose the restrictions under H_0, we are left with the restricted model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u. \quad [4.36]$$

In this subsection, we assume that both the unrestricted and restricted models contain an intercept, because that is the case most widely encountered in practice.

Now, for the test statistic itself. Earlier, we suggested that looking at the relative increase in the SSR when moving from the unrestricted to the restricted model should be informative for testing the hypothesis (4.35). The ***F* statistic** (or *F ratio*) is defined by

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)}, \quad [4.37]$$

where SSR_r is the sum of squared residuals from the restricted model and SSR_{ur} is the sum of squared residuals from the unrestricted model.

You should immediately notice that, because SSR_r can be no smaller than SSR_{ur}, the F statistic is *always* nonnegative (and almost always strictly positive). Thus, if you compute a negative F statistic, then something is wrong; the order of the SSRs in the numerator of F has usually been reversed. Also, the SSR in the denominator of F is the SSR from the *unrestricted* model. The easiest way to remember where the SSRs appear is to think of F as measuring the relative increase in SSR when moving from the unrestricted to the restricted model.

The difference in SSRs in the numerator of F is divided by q, which is the number of restrictions imposed in moving from the unrestricted to the restricted model (q independent variables are dropped). Therefore, we can write

$$q = \textbf{numerator degrees of freedom} = df_r - df_{ur}, \quad [4.38]$$

which also shows that q is the difference in degrees of freedom between the restricted and unrestricted models. (Recall that df = number of observations − number of estimated parameters.)

GOING FURTHER 4.4

Consider relating individual performance on a standardized test, *score*, to a variety of other variables. School factors include average class size, per-student expenditures, average teacher compensation, and total school enrollment. Other variables specific to the student are family income, mother's education, father's education, and number of siblings. The model is

$$score = \beta_0 + \beta_1 classize + \beta_2 expend + \beta_3 tchcomp + \beta_4 enroll + \beta_5 faminc + \beta_6 motheduc + \beta_7 fatheduc + \beta_8 siblings + u.$$

State the null hypothesis that student-specific variables have no effect on standardized test performance once school-related factors have been controlled for. What are k and q for this example? Write down the restricted version of the model.

Because the restricted model has fewer parameters—and each model is estimated using the same n observations—df_r is always greater than df_{ur}.

The SSR in the denominator of F is divided by the degrees of freedom in the unrestricted model:

$$n - k - 1 = \textbf{denominator degrees of freedom} = df_{ur}. \quad [4.39]$$

In fact, the denominator of F is just the unbiased estimator of $\sigma^2 = \text{Var}(u)$ in the unrestricted model.

In a particular application, computing the F statistic is easier than wading through the somewhat cumbersome notation used to describe the general case. We first obtain the degrees of freedom in the unrestricted model, df_{ur}. Then, we count how many variables are excluded in the restricted model; this is q. The SSRs are reported with every OLS regression, and so forming the F statistic is simple.

In the major league baseball salary regression, $n = 353$, and the full model (4.28) contains six parameters. Thus, $n - k - 1 = df_{ur} = 353 - 6 = 347$. The restricted model (4.32) contains three fewer independent variables than (4.28), and so $q = 3$. Thus, we have all of the ingredients to compute the F statistic; we hold off doing so until we know what to do with it.

To use the F statistic, we must know its sampling distribution under the null in order to choose critical values and rejection rules. It can be shown that, under H_0 (and assuming the CLM assumptions hold), F is distributed as an F random variable with $(q, n - k - 1)$ degrees of freedom. We write this as

$$F \sim F_{q,n-k-1}.$$

The distribution of $F_{q,n-k-1}$ is readily tabulated and available in statistical tables (see Table G.3) and, even more importantly, in statistical software.

We will not derive the F distribution because the mathematics is very involved. Basically, it can be shown that equation (4.37) is actually the ratio of two independent chi-square random variables, divided by their respective degrees of freedom. The numerator chi-square random variable has q degrees of freedom, and the chi-square in the denominator has $n - k - 1$ degrees of freedom. This is the definition of an F distributed random variable (see Math Refresher B).

It is pretty clear from the definition of F that we will reject H_0 in favor of H_1 when F is sufficiently "large." How large depends on our chosen significance level. Suppose that we have decided on a 5% level test. Let c be the 95$^{\text{th}}$ percentile in the $F_{q,n-k-1}$ distribution. This critical value depends on q (the numerator df) and $n - k - 1$ (the denominator df). It is important to keep the numerator and denominator degrees of freedom straight.

The 10%, 5%, and 1% critical values for the F distribution are given in Table G.3. The rejection rule is simple. Once c has been obtained, we reject H_0 in favor of H_1 at the chosen significance level if

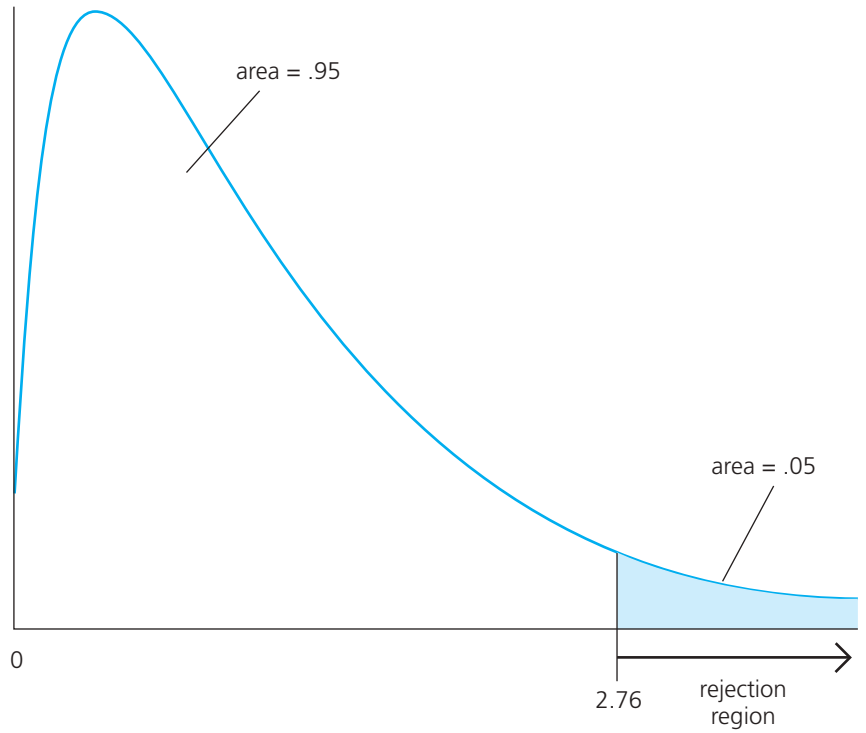
$$F > c. \quad [4.40]$$

With a 5% significance level, $q = 3$, and $n - k - 1 = 60$, the critical value is $c = 2.76$. We would reject H_0 at the 5% level if the computed value of the F statistic exceeds 2.76. The 5% critical value and rejection region are shown in Figure 4.7. For the same degrees of freedom, the 1% critical value is 4.13.

In most applications, the numerator degrees of freedom (q) will be notably smaller than the denominator degrees of freedom $(n - k - 1)$. Applications where $n - k - 1$ is small are unlikely to be successful because the parameters in the unrestricted model will probably not be precisely estimated. When the denominator df reaches about 120, the F distribution is no longer sensitive to it. (This is entirely analogous to the t distribution being well approximated by the standard normal distribution as the df gets large.) Thus, there is an entry in the table for the denominator $df = \infty$, and this is what we use with large samples (because $n - k - 1$ is then large). A similar statement holds for a very large numerator df, but this rarely occurs in applications.

If H_0 is rejected, then we say that $x_{k-q+1}, \ldots, x_k$ are **jointly statistically significant** (or just *jointly significant*) at the appropriate significance level. This test alone does not allow us to say which of the variables has a partial effect on y; they may all affect y or maybe only one affects y.

Figure 4.7 The 5% critical value and rejection region in an $F_{3,60}$ distribution.



If the null is not rejected, then the variables are **jointly insignificant**, which often justifies dropping them from the model.

For the major league baseball example with three numerator degrees of freedom and 347 denominator degrees of freedom, the 5% critical value is 2.60, and the 1% critical value is 3.78. We reject H_0 at the 1% level if F is above 3.78; we reject at the 5% level if F is above 2.60.

We are now in a position to test the hypothesis that we began this section with: after controlling for *years* and *gamesyr*, the variables *bavg*, *hrunsyr*, and *rbisyr* have no effect on players' salaries. In practice, it is easiest to first compute $(SSR_r - SSR_{ur})/SSR_{ur}$ and to multiply the result by $(n - k - 1)/q$; the reason the formula is stated as in (4.37) is that it makes it easier to keep the numerator and denominator degrees of freedom straight. Using the SSRs in (4.31) and (4.33), we have

$$F = \frac{(198.311 - 183.186)}{183.186} \cdot \frac{347}{3} \approx 9.55.$$

This number is well above the 1% critical value in the F distribution with 3 and 347 degrees of freedom, and so we soundly reject the hypothesis that *bavg*, *hrunsyr*, and *rbisyr* have no effect on salary.

The outcome of the joint test may seem surprising in light of the insignificant t statistics for the three variables. What is happening is that the two variables *hrunsyr* and *rbisyr* are highly correlated, and this multicollinearity makes it difficult to uncover the partial effect of each variable; this is reflected in the individual t statistics. The F statistic tests whether these variables (including *bavg*) are *jointly* significant, and multicollinearity between *hrunsyr* and *rbisyr* is much less relevant for testing this hypothesis. In Computer Exercise C5, you are asked to reestimate the model while dropping

rbisyr, in which case *hrunsyr* becomes very significant. The same is true for *rbisyr* when *hrunsyr* is dropped from the model.

The F statistic is often useful for testing exclusion of a group of variables when the variables in the group are highly correlated. For example, suppose we want to test whether firm performance affects the salaries of chief executive officers. There are many ways to measure firm performance, and it probably would not be clear ahead of time which measures would be most important. Because measures of firm performance are likely to be highly correlated, hoping to find individually significant measures might be asking too much due to multicollinearity. But an F test can be used to determine whether, as a group, the firm performance variables affect salary.

4-5b Relationship between F and t Statistics

We have seen in this section how the F statistic can be used to test whether a group of variables should be included in a model. What happens if we apply the F statistic to the case of testing significance of a *single* independent variable? This case is certainly not ruled out by the previous development. For example, we can take the null to be H_0: $\beta_k = 0$ and $q = 1$ (to test the single exclusion restriction that x_k can be excluded from the model). From Section 4-2, we know that the t statistic on β_k can be used to test this hypothesis. The question, then, is: do we have two separate ways of testing hypotheses about a single coefficient? The answer is no. It can be shown that the F statistic for testing exclusion of a single variable is equal to the *square* of the corresponding t statistic. Because t^2_{n-k-1} has an $F_{1,n-k-1}$ distribution, the two approaches lead to exactly the same outcome, provided that the alternative is two-sided. The t statistic is more flexible for testing a single hypothesis because it can be directly used to test against one-sided alternatives. Because t statistics are also easier to obtain than F statistics, there is really no reason to use an F statistic to test hypotheses about a single parameter.

We have already seen in the salary regressions for major league baseball players that two (or more) variables that each have insignificant t statistics can be jointly very significant. It is also possible that, in a group of several explanatory variables, one variable has a significant t statistic but the group of variables is jointly insignificant at the usual significance levels. What should we make of this kind of outcome? For concreteness, suppose that in a model with many explanatory variables we cannot reject the null hypothesis that β_1, β_2, β_3, β_4, and β_5 are all equal to zero at the 5% level, yet the t statistic for $\hat{\beta}_1$ is significant at the 5% level. Logically, we cannot have $\beta_1 \neq 0$ but also have β_1, β_2, β_3, β_4, and β_5 all equal to zero! But as a matter of testing, it is possible that we can group a bunch of insignificant variables with a significant variable and conclude that the entire set of variables is jointly insignificant. (Such possible conflicts between a t test and a joint F test give another example of why we should not "accept" null hypotheses; we should only fail to reject them.) The F statistic is intended to detect whether a set of coefficients is different from zero, but it is never the best test for determining whether a single coefficient is different from zero. The t test is best suited for testing a single hypothesis. (In statistical terms, an F statistic for joint restrictions including $\beta_1 = 0$ will have less power for detecting $\beta_1 \neq 0$ than the usual t statistic. See Section C-6 in Math Refresher C for a discussion of the power of a test.)

Unfortunately, the fact that we can sometimes hide a statistically significant variable along with some insignificant variables could lead to abuse if regression results are not carefully reported. For example, suppose that, in a study of the determinants of loan-acceptance rates at the city level, x_1 is the fraction of black households in the city. Suppose that the variables x_2, x_3, x_4, and x_5 are the fractions of households headed by different age groups. In explaining loan rates, we would include measures of income, wealth, credit ratings, and so on. Suppose that age of household head has no effect on loan approval rates, once other variables are controlled for. Even if race has a marginally significant effect, it is possible that the race and age variables could be jointly insignificant. Someone wanting to conclude that race is not a factor could simply report something like "Race and age variables were added to the equation, but they were jointly insignificant at the 5% level."

Hopefully, peer review prevents these kinds of misleading conclusions, but you should be aware that such outcomes are possible.

Often, when a variable is very statistically significant and it is tested jointly with another set of variables, the set will be jointly significant. In such cases, there is no logical inconsistency in rejecting both null hypotheses.

4-5c The *R*-Squared Form of the *F* Statistic

For testing exclusion restrictions, it is often more convenient to have a form of the F statistic that can be computed using the R-squareds from the restricted and unrestricted models. One reason for this is that the R-squared is always between zero and one, whereas the SSRs can be very large depending on the unit of measurement of y, making the calculation based on the SSRs tedious. Using the fact that $SSR_r = SST(1 - R_r^2)$ and $SSR_{ur} = SST(1 - R_{ur}^2)$, we can substitute into (4.37) to obtain

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)} = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/df_{ur}} \qquad [4.41]$$

(note that the SST terms cancel everywhere). This is called the ***R*-squared form of the *F* statistic**. [At this point, you should be cautioned that although equation (4.41) is very convenient for testing exclusion restrictions, it cannot be applied for testing all linear restrictions. As we will see when we discuss testing general linear restrictions, the sum of squared residuals form of the F statistic is sometimes needed.]

Because the R-squared is reported with almost all regressions (whereas the SSR is not), it is easy to use the R-squareds from the unrestricted and restricted models to test for exclusion of some variables. Particular attention should be paid to the order of the R-squareds in the numerator: the *unrestricted* R-squared comes first [contrast this with the SSRs in (4.37)]. Because $R_{ur}^2 > R_r^2$, this shows again that F will always be positive.

In using the R-squared form of the test for excluding a set of variables, it is important to *not* square the R-squared before plugging it into formula (4.41); the squaring has already been done. All regressions report R^2, and these numbers are plugged directly into (4.41). For the baseball salary example, we can use (4.41) to obtain the F statistic:

$$F = \frac{(.6278 - .5971)}{(1 - .6278)} \cdot \frac{347}{3} \approx 9.54,$$

which is very close to what we obtained before. (The difference is due to rounding error.)

EXAMPLE 4.9 **Parents' Education in a Birth Weight Equation**

As another example of computing an F statistic, consider the following model to explain child birth weight in terms of various factors:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + u, \qquad [4.42]$$

where

bwght = birth weight, in pounds.
cigs = average number of cigarettes the mother smoked per day during pregnancy.
parity = the birth order of this child.
faminc = annual family income.
motheduc = years of schooling for the mother.
fatheduc = years of schooling for the father.

Let us test the null hypothesis that, after controlling for *cigs*, *parity*, and *faminc*, parents' education has no effect on birth weight. This is stated as H_0: $\beta_4 = 0, \beta_5 = 0$, and so there are $q = 2$ exclusion restrictions to be tested. There are $k + 1 = 6$ parameters in the unrestricted model (4.42); so the *df* in the unrestricted model is $n - 6$, where n is the sample size.

We will test this hypothesis using the data in BWGHT. This data set contains information on 1,388 births, but we must be careful in counting the observations used in testing the null hypothesis. It turns out that information on at least one of the variables *motheduc* and *fatheduc* is missing for 197 births in the sample; these observations cannot be included when estimating the unrestricted model. Thus, we really have $n = 1{,}191$ observations, and so there are $1{,}191 - 6 = 1{,}185$ *df* in the unrestricted model. We must be sure to use these *same* 1,191 observations when estimating the restricted model (not the full 1,388 observations that are available). Generally, when estimating the restricted model to compute an *F* test, we must use the same observations to estimate the unrestricted model; otherwise, the test is not valid. When there are no missing data, this will not be an issue.

The numerator *df* is 2, and the denominator *df* is 1,185; from Table G.3, the 5% critical value is $c = 3.0$. Rather than report the complete results, for brevity, we present only the *R*-squareds. The *R*-squared for the full model turns out to be $R^2_{ur} = .0387$. When *motheduc* and *fatheduc* are dropped from the regression, the *R*-squared falls to $R^2_r = .0364$. Thus, the *F* statistic is $F = [(.0387 - .0364)/(1 - .0387)](1{,}185/2) = 1.42$; because this is well below the 5% critical value, we fail to reject H_0. In other words, *motheduc* and *fatheduc* are jointly insignificant in the birth weight equation. Most statistical packages have built-in commands for testing multiple hypotheses after OLS estimation, and so one need not worry about making the mistake of running the two regressions on different data sets. Typically, the commands are applied after estimation of the unrestricted model, which means the smaller subset of data is used whenever there are missing values on some variables. Formulas for computing the *F* statistic using matrix algebra—see Advanced Treatment E—do not require estimation of the restricted model.

4-5d Computing *p*-Values for *F* Tests

For reporting the outcomes of *F* tests, *p*-values are especially useful. Because the *F* distribution depends on the numerator and denominator *df*, it is difficult to get a feel for how strong or weak the evidence is against the null hypothesis simply by looking at the value of the *F* statistic and one or two critical values.

In the *F* testing context, the *p*-value is defined as

$$p\text{-value} = P(\mathcal{F} > F), \qquad [4.43]$$

where, for emphasis, we let $\mathcal{F}$ denote an *F* random variable with $(q, n - k - 1)$ degrees of freedom, and *F* is the actual value of the test statistic. The *p*-value still has the same interpretation as it did for *t* statistics: it is the probability of observing a value of *F* at least as large as we did, *given* that the null hypothesis is true. A small *p*-value is evidence against H_0. For example, *p*-value = .016 means that the chance of observing a value of *F* as large as we did when the null hypothesis was true is only 1.6%; we usually reject H_0 in such cases. If the *p*-value = .314, then the chance of observing a value of the *F* statistic as large as we did under the null hypothesis is 31.4%. Most would find this to be pretty weak evidence against H_0.

GOING FURTHER 4.5

The data in ATTEND were used to estimate the two equations

$$\widehat{atndrte} = 47.13 + 13.37\ priGPA$$
$$(2.87)\quad (1.09)$$
$$n = 680,\ R^2 = .183$$

and

$$\widehat{atndrte} = 75.70 + 17.26\ priGPA - 1.72\ ACT$$
$$(3.88)\quad (1.08)\qquad (?)$$
$$n = 680,\ R^2 = .291,$$

where, as always, standard errors are in parentheses; the standard error for *ACT* is missing in the second equation. What is the *t* statistic for the coefficient on *ACT*? (*Hint:* First compute the *F* statistic for significance of *ACT*.)

As with t testing, once the p-value has been computed, the F test can be carried out at any significance level. For example, if the p-value $= .024$, we reject H_0 at the 5% significance level but not at the 1% level.

The p-value for the F test in Example 4.9 is .238, and so the null hypothesis that $\beta_{motheduc}$ and $\beta_{fatheduc}$ are both zero is not rejected at even the 20% significance level.

Many econometrics packages have a built-in feature for testing multiple exclusion restrictions. These packages have several advantages over calculating the statistics by hand: we will less likely make a mistake, p-values are computed automatically, and the problem of missing data, as in Example 4.9, is handled without any additional work on our part.

4-5e The *F* Statistic for Overall Significance of a Regression

A special set of exclusion restrictions is routinely tested by most regression packages. These restrictions have the same interpretation, regardless of the model. In the model with k independent variables, we can write the null hypothesis as

$$H_0: x_1, x_2, \ldots, x_k \text{ do not help to explain } y.$$

This null hypothesis is, in a way, very pessimistic. It states that *none* of the explanatory variables has an effect on y. Stated in terms of the parameters, the null is that all slope parameters are zero:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0, \qquad [4.44]$$

and the alternative is that at least one of the β_j is different from zero. Another useful way of stating the null is that $H_0: E(y|x_1, x_2, \ldots, x_k) = E(y)$, so that knowing the values of $x_1, x_2, \ldots, x_k$ does not affect the expected value of y.

There are k restrictions in (4.44), and when we impose them, we get the restricted model

$$y = \beta_0 + u; \qquad [4.45]$$

all independent variables have been dropped from the equation. Now, the R-squared from estimating (4.45) is zero; none of the variation in y is being explained because there are no explanatory variables. Therefore, the F statistic for testing (4.44) can be written as

$$\frac{R^2/k}{(1 - R^2)/(n - k - 1)}, \qquad [4.46]$$

where R^2 is just the usual R-squared from the regression of y on $x_1, x_2, \ldots, x_k$.

Most regression packages report the F statistic in (4.46) automatically, which makes it tempting to use this statistic to test general exclusion restrictions. You must avoid this temptation. The F statistic in (4.41) is used for general exclusion restrictions; it depends on the R-squareds from the restricted and unrestricted models. The special form of (4.46) is valid only for testing joint exclusion of *all* independent variables. This is sometimes called determining the **overall significance of the regression**.

If we fail to reject (4.44), then there is no evidence that any of the independent variables help to explain y. This usually means that we must look for other variables to explain y. For Example 4.9, the F statistic for testing (4.44) is about 9.55 with $k = 5$ and $n - k - 1 = 1{,}185$ df. The p-value is zero to four places after the decimal point, so that (4.44) is rejected very strongly. Thus, we conclude that the variables in the *bwght* equation *do* explain some variation in *bwght*. The amount explained is not large: only 3.87%. But the seemingly small R-squared results in a highly significant F statistic. That is why we must compute the F statistic to test for joint significance and not just look at the size of the R-squared.

Occasionally, the F statistic for the hypothesis that all independent variables are jointly insignificant is the focus of a study. Problem 10 asks you to use stock return data to test whether stock returns over a four-year horizon are predictable based on information known only at the beginning of the period. Under the *efficient markets hypothesis*, the returns should not be predictable; the null hypothesis is precisely (4.44).

4-5f Testing General Linear Restrictions

Testing exclusion restrictions is by far the most important application of F statistics. Sometimes, however, the restrictions implied by a theory are more complicated than just excluding some independent variables. It is still straightforward to use the F statistic for testing.

As an example, consider the following equation:

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u, \quad [4.47]$$

where

$price$ = house price.
$assess$ = the assessed housing value(before the house was sold).
$lotsize$ = size of the lot, in square feet.
$sqrft$ = square footage.
$bdrms$ = number of bedrooms.

Now, suppose we would like to test whether the assessed housing price is a rational valuation. If this is the case, then a 1% change in *assess* should be associated with a 1% change in *price*; that is, $\beta_1 = 1$. In addition, *lotsize*, *sqrft*, and *bdrms* should not help to explain log(*price*), once the assessed value has been controlled for. Together, these hypotheses can be stated as

$$H_0: \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0. \quad [4.48]$$

Four restrictions have to be tested; three are exclusion restrictions, but $\beta_1 = 1$ is not. How can we test this hypothesis using the F statistic?

As in the exclusion restriction case, we estimate the unrestricted model, (4.47) in this case, and then impose the restrictions in (4.48) to obtain the restricted model. It is the second step that can be a little tricky. But all we do is plug in the restrictions. If we write (4.47) as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u, \quad [4.49]$$

then the restricted model is $y = \beta_0 + x_1 + u$. Now, to impose the restriction that the coefficient on x_1 is unity, we must estimate the following model:

$$y - x_1 = \beta_0 + u. \quad [4.50]$$

This is just a model with an intercept (β_0) but with a different dependent variable than in (4.49). The procedure for computing the F statistic is the same: estimate (4.50), obtain the $SSR(SSR_r)$, and use this with the unrestricted SSR from (4.49) in the F statistic (4.37). We are testing $q = 4$ restrictions, and there are $n - 5$ *df* in the unrestricted model. The F statistic is simply $[(SSR_r - SSR_{ur})/SSR_{ur}][(n - 5)/4]$.

Before illustrating this test using a data set, we must emphasize one point: we cannot use the R-squared form of the F statistic for this example because the dependent variable in (4.50) is different from the one in (4.49). This means the total sum of squares from the two regressions will be different, and (4.41) is no longer equivalent to (4.37). As a general rule, the SSR form of the F statistic should be used if a different dependent variable is needed in running the restricted regression.

The estimated unrestricted model using the data in HPRICE1 is

$$\widehat{\log(price)} = \underset{(.570)}{.264} + \underset{(.151)}{1.043} \log(assess) + \underset{(.0386)}{.0074} \log(lotsize) - \underset{(.1384)}{.1032} \log(sqrft) + \underset{(.0221)}{.0338} bdrms$$
$$n = 88, SSR = 1.822, R^2 = .773.$$

If we use separate t statistics to test each hypothesis in (4.48), we fail to reject each one. But rationality of the assessment is a joint hypothesis, so we should test the restrictions jointly. The SSR from the restricted model turns out to be $SSR_r = 1.880$, and so the F statistic is $[(1.880 - 1.822)/1.822](83/4) = .661$. The 5% critical value in an F distribution with (4,83) df is about 2.50, and so we fail to reject H_0. There is essentially no evidence against the hypothesis that the assessed values are rational.

4-6 Reporting Regression Results

We end this chapter by providing a few guidelines on how to report multiple regression results for relatively complicated empirical projects. This should help you to read published works in the applied social sciences, while also preparing you to write your own empirical papers. We will expand on this topic in the remainder of the text by reporting results from various examples, but many of the key points can be made now.

Naturally, the estimated OLS coefficients should always be reported. For the key variables in an analysis, you should *interpret* the estimated coefficients (which often requires knowing the units of measurement of the variables). For example, is an estimate an elasticity, or does it have some other interpretation that needs explanation? The economic or practical importance of the estimates of the key variables should be discussed.

The standard errors should always be included along with the estimated coefficients. Some authors prefer to report the t statistics rather than the standard errors (and sometimes just the absolute value of the t statistics). Although nothing is really wrong with this, there is some preference for reporting standard errors. First, it forces us to think carefully about the null hypothesis being tested; the null is not always that the population parameter is zero. Second, having standard errors makes it easier to compute confidence intervals.

The R-squared from the regression should always be included. We have seen that, in addition to providing a goodness-of-fit measure, it makes calculation of F statistics for exclusion restrictions simple. Reporting the sum of squared residuals and the standard error of the regression is sometimes a good idea, but it is not crucial. The number of observations used in estimating any equation should appear near the estimated equation.

If only a couple of models are being estimated, the results can be summarized in equation form, as we have done up to this point. However, in many papers, several equations are estimated with many different sets of independent variables. We may estimate the same equation for different groups of people, or even have equations explaining different dependent variables. In such cases, it is better to summarize the results in one or more tables. The dependent variable should be indicated clearly in the table, and the independent variables should be listed in the first column. Standard errors (or t statistics) can be put in parentheses below the estimates.

EXAMPLE 4.10 Salary-Pension Tradeoff for Teachers

Let *totcomp* denote average total annual compensation for a teacher, including salary and all fringe benefits (pension, health insurance, and so on). Extending the standard wage equation, total compensation should be a function of productivity and perhaps other characteristics. As is standard, we use logarithmic form:

$$\log(totcomp) = f(productivity\ characteristics, other\ factors),$$

where $f(\cdot)$ is some function (unspecified for now). Write

$$totcomp = salary + benefits = salary\left(1 + \frac{benefits}{salary}\right).$$

This equation shows that total compensation is the product of two terms: *salary* and $1 + b/s$, where b/s is shorthand for the "benefits to salary ratio." Taking the log of this equation gives $\log(totcomp) = \log(salary) + \log(1 + b/s)$. Now, for "small" b/s, $\log(1 + b/s) \approx b/s$; we will use this approximation. This leads to the econometric model

$$\log(salary) = \beta_0 + \beta_1(b/s) + other\ factors.$$

Testing the salary-benefits tradeoff then is the same as a test of H_0: $\beta_1 = -1$ against H_1: $\beta_1 \neq -1$.

We use the data in MEAP93 to test this hypothesis. These data are averaged at the school level, and we do not observe very many other factors that could affect total compensation. We will include controls for size of the school (*enroll*), staff per thousand students (*staff*), and measures such as the school dropout and graduation rates. The average b/s in the sample is about .205, and the largest value is .450.

The estimated equations are given in Table 4.1, where standard errors are given in parentheses below the coefficient estimates. The key variable is b/s, the benefits-salary ratio.

From the first column in Table 4.1, we see that, without controlling for any other factors, the OLS coefficient for b/s is $-.825$. The t statistic for testing the null hypothesis H_0: $\beta_1 = -1$ is $t = (-.825 + 1)/.200 = .875$, and so the simple regression fails to reject H_0. After adding controls for school size and staff size (which roughly captures the number of students taught by each teacher), the estimate of the b/s coefficient becomes $-.605$. Now, the test of $\beta_1 = -1$ gives a t statistic of about 2.39; thus, H_0 is rejected at the 5% level against a two-sided alternative. The variables log(*enroll*) and log(*staff*) are very statistically significant.

GOING FURTHER 4.6

How does adding *droprate* and *gradrate* affect the estimate of the salary-benefits tradeoff? Are these variables jointly significant at the 5% level? What about the 10% level?

TABLE 4.1 Testing the Salary-Benefits Tradeoff

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
b/s	–.825 (.200)	–.605 (.165)	–.589 (.165)
log(*enroll*)	———	.0874 (.0073)	.0881 (.0073)
log(*staff*)	———	–.222 (.050)	–.218 (.050)
droprate	———	———	–.00028 (.00161)
gradrate	———	———	.00097 .00066)
intercept	10.523 (0.042)	10.884 (0.252)	10.738 (0.258)
Observations *R*-squared	408 .040	408 .353	408 .361

4-7 Revisiting Causal Effects and Policy Analysis

In Section 3-7e we showed how multiple regression can be used to obtain unbiased estimators of causal, or treatment, effects in the context of policy interventions, provided we have controls sufficient to ensure that participation assignment is unconfounded. In particular, with a constant treatment effect, τ, we derived

$$\mathrm{E}(y|w, \mathbf{x}) = \alpha + \tau w + \mathbf{x}\boldsymbol{\gamma} = \alpha + \tau w + \gamma_1 x_1 + \cdots + \gamma_k x_k,$$

where y is the outcome or response, w is the binary policy (treatment) variable, and the x_j are the controls that account for nonrandom assignment. We know that the OLS estimator of τ is unbiased because MLR.1 and MLR.4 hold (and we have random sampling from the population). If we add MLR.5 and MLR.6, we can perform exact inference on τ. For example, the null hypothesis of no policy effect is $\mathrm{H}_0: \tau = 0$, and we can test this hypothesis—against a one-sided or two-sided alternative—using a standard t statistic. Regardless of the magnitude of the estimate $\hat{\tau}$, most researchers and administrators will not be convinced that an intervention or policy is effective unless $\hat{\tau}$ is statistically different from zero (and with the expected sign) at a sufficiently small significance level. As in any context, it is important to discuss the sign and magnitude of $\hat{\tau}$ in addition to its statistical significance. Probably of more interest is to obtain a 95% confidence interval for τ, which gives us a plausible range of values for the population treatment effect.

We can also test hypotheses about the γ_j, but, in a policy environment, we are rarely concerned about the statistical significance of the x_j except perhaps as a logical check on the regression results. For example, we should expect past labor market earnings to positively predict current labor market earnings.

We now revisit Example 3.7, which contains the estimated effect of a job training program using JTRAIN98.

EXAMPLE 4.11 Evaluating a Job Training Program

We reproduce the simple and multiple regression estimates and now put the standard errors below the coefficients. Recall that the outcome variable, *earn*98, is measured in thousands of dollars:

$$\widehat{earn98} = 10.61 - 2.05\ train \qquad [4.51]$$
$$(0.28)\quad (0.48)$$
$$n = 1{,}130,\ R^2 = 0.016$$

$$\widehat{earn98} = 4.67 + 2.41\ train + .373\ earn96 + .363\ educ - .181\ age + 2.48\ married \qquad [4.52]$$
$$(1.15)\quad (0.44)\qquad (.019)\qquad (.064)\qquad (.019)\qquad (0.43)$$
$$n = 1{,}130,\ R^2 = 0.405$$

As discussed in Example 3.7, the change in the sign of coefficient on *train* is striking when moving from simple to multiple regression. Moreover, the t statistic in (4.51) is $-2.05/0.48 \approx -4.27$, which gives a very statistically significant and practically large *negative* effect of the program. By contrast, the t statistic in (4.52) is about 5.47, which shows a strongly statistically significant and *positive* effect. It is pretty clear that we prefer the multiple regression results for evaluating the job training program. Of course, it could be that we have omitted some important controls in (4.52), but at a minimum we know that we can account for some important differences across workers.

Perhaps now is a good time to revisit the the multicollinearity issue, which we raised in Section 3-4a. Recall that collinearity arises only in the context of multiple regression, and so our discussion is relevant only for equation (4.52). In this equation, it could be that two or more of the control variables in (4.52) are highly correlated; or maybe not. The point is that we do not care. The reason we

include *earn*96, *educ, age*, and *married* is to control for differences in men that at least partly determine participation in the job training program, hopefully leading to an unbiased estimator of the treatment effect. We are not worried about how well we estimate the coefficients on the control variables, and including highly correlated variables among the x_j has nothing to do with obtaining a reliable estimate of τ. Computer Exercise C14 asks you to add a binary variable for whether the man was unemployed in 1996, which is strongly related to earnings in 1996: *unem*96 = 0 means *earn*96 = 0. And yet, correlation between *unem*96 and *earn*96 is of essentially no concern. If we could observe earnings in 1995, *earn*95, we would likely include it, too, even though it is likely to be highly correlated with *earn*96.

Summary

In this chapter, we have covered the very important topic of statistical inference, which allows us to infer something about the population model from a random sample. We summarize the main points:

1. Under the classical linear model assumptions MLR.1 through MLR.6, the OLS estimators are normally distributed.
2. Under the CLM assumptions, the *t* statistics have *t* distributions under the null hypothesis.
3. We use *t* statistics to test hypotheses about a single parameter against one- or two-sided alternatives, using one- or two-tailed tests, respectively. The most common null hypothesis is H_0: $\beta_j = 0$, but we sometimes want to test other values of β_j under H_0.
4. In classical hypothesis testing, we first choose a significance level, which, along with the *df* and alternative hypothesis, determines the critical value against which we compare the *t* statistic. It is more informative to compute the *p*-value for a *t* test—the smallest significance level for which the null hypothesis is rejected—so that the hypothesis can be tested at any significance level.
5. Under the CLM assumptions, confidence intervals can be constructed for each β_j. These CIs can be used to test any null hypothesis concerning β_j against a two-sided alternative.
6. Single hypothesis tests concerning more than one β_j can always be tested by rewriting the model to contain the parameter of interest. Then, a standard *t* statistic can be used.
7. The *F* statistic is used to test multiple exclusion restrictions, and there are two equivalent forms of the test. One is based on the SSRs from the restricted and unrestricted models. A more convenient form is based on the *R*-squareds from the two models.
8. When computing an *F* statistic, the numerator *df* is the number of restrictions being tested, while the denominator *df* is the degrees of freedom in the unrestricted model.
9. The alternative for *F* testing is two-sided. In the classical approach, we specify a significance level which, along with the numerator *df* and the denominator *df*, determines the critical value. The null hypothesis is rejected when the statistic, *F*, exceeds the critical value, *c*. Alternatively, we can compute a *p*-value to summarize the evidence against H_0.
10. General multiple linear restrictions can be tested using the sum of squared residuals form of the *F* statistic.
11. The *F* statistic for the overall significance of a regression tests the null hypothesis that *all* slope parameters are zero, with the intercept unrestricted. Under H_0, the explanatory variables have no effect on the expected value of *y*.
12. When data are missing on one or more explanatory variables, one must be careful when computing *F* statistics "by hand," that is, using either the sum of squared residuals or *R*-squareds from the two regressions. Whenever possible it is best to leave the calculations to statistical packages that have built-in commands, which work with or without missing data.
13. Statistical inference is important for program evaluation and policy analysis. Rarely is it enough to report only the economic (or practical) significance of our estimates. We, and others, must be convinced that moderate to large estimates of treatment effects are not due purely to sampling variation. Obtaining a *p*-value for the null hypothesis that the effect is zero, or, even better, obtaining a 95% confidence interval, allows us to determine statistic significance in addition to economic significance.

THE CLASSICAL LINEAR MODEL ASSUMPTIONS

Now is a good time to review the full set of classical linear model (CLM) assumptions for cross-sectional regression. Following each assumption is a comment about its role in multiple regression analysis.

Assumption MLR.1 (Linear in Parameters)

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u,$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term.

Assumption MLR.1 describes the population relationship we hope to estimate, and explicitly sets out the β_j—the ceteris paribus population effects of the x_j on y—as the parameters of interest.

Assumption MLR.2 (Random Sampling)

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, \ldots, n\}$, following the population model in Assumption MLR.1.

This random sampling assumption means that we have data that can be used to estimate the β_j, and that the data have been chosen to be representative of the population described in Assumption MLR.1.

Assumption MLR.3 (No Perfect Collinearity)

In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact *linear* relationships among the independent variables.

Once we have a sample of data, we need to know that we can use the data to compute the OLS estimates, the $\hat{\beta}_j$. This is the role of Assumption MLR.3: if we have sample variation in each independent variable and no exact linear relationships among the independent variables, we can compute the $\hat{\beta}_j$.

Assumption MLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any values of the explanatory variables. In other words, $\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$.

As we discussed in the text, assuming that the unobserved factors are, on average, unrelated to the explanatory variables is key to deriving the first statistical property of each OLS estimator: its unbiasedness for the corresponding population parameter. Of course, all of the previous assumptions are used to show unbiasedness.

Assumption MLR.5 (Homoskedasticity)

The error u has the same variance given any values of the explanatory variables. In other words,

$$\mathrm{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2.$$

Compared with Assumption MLR.4, the homoskedasticity assumption is of secondary importance; in particular, Assumption MLR.5 has no bearing on the unbiasedness of the $\hat{\beta}_j$. Still, homoskedasticity has two important implications: (1) We can derive formulas for the sampling variances whose components are easy to characterize; (2) We can conclude, under the Gauss-Markov assumptions MLR.1 through MLR.5, that the OLS estimators have smallest variance among *all* linear, unbiased estimators.

Assumption MLR.6 (Normality)

The population error u is *independent* of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \mathrm{Normal}(0, \sigma^2)$.

In this chapter, we added Assumption MLR.6 to obtain the exact sampling distributions of t statistics and F statistics, so that we can carry out exact hypotheses tests. In the next chapter, we will see that MLR.6 can be dropped if we have a reasonably large sample size. Assumption MLR.6 does imply a stronger efficiency property of OLS: the OLS estimators have smallest variance among *all* unbiased estimators; the comparison group is no longer restricted to estimators linear in the $\{y_i: i = 1, 2, \ldots, n\}$.

Key Terms

Alternative Hypothesis
Classical Linear Model
Classical Linear Model (CLM) Assumptions
Confidence Interval (CI)
Critical Value
Denominator Degrees of Freedom
Economic Significance
Exclusion Restrictions
F Statistic
Joint Hypotheses Test
Jointly Insignificant
Jointly Statistically Significant
Minimum Variance Unbiased Estimators
Multiple Hypotheses Test
Multiple Restrictions
Normality Assumption
Null Hypothesis
Numerator Degrees of Freedom
One-Sided Alternative
One-Tailed Test
Overall Significance of the Regression
p-Value
Practical Significance
R-squared Form of the *F* Statistic
Rejection Rule
Restricted Model
Significance Level
Statistically Insignificant
Statistically Significant
t Ratio
t Statistic
Two-Sided Alternative
Two-Tailed Test
Unrestricted Model

Problems

1 Which of the following can cause the usual OLS *t* statistics to be invalid (that is, not to have *t* distributions under H_0)?

(i) Heteroskedasticity.

(ii) A sample correlation coefficient of .95 between two independent variables that are in the model.

(iii) Omitting an important explanatory variable.

2 Consider an equation to explain salaries of CEOs in terms of annual firm sales, return on equity (*roe*, in percentage form), and return on the firm's stock (*ros*, in percentage form):

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 ros + u.$$

(i) In terms of the model parameters, state the null hypothesis that, after controlling for *sales* and *roe*, *ros* has no effect on CEO salary. State the alternative that better stock market performance increases a CEO's salary.

(ii) Using the data in CEOSAL1, the following equation was obtained by OLS:

$$\widehat{\log(salary)} = 4.32 + .280 \log(sales) + .0174\, roe + .00024\, ros$$
$$(.32)\ (.035)\qquad (.0041)\qquad (.00054)$$
$$n = 209,\ R^2 = .283.$$

By what percentage is *salary* predicted to increase if *ros* increases by 50 points? Does *ros* have a practically large effect on *salary*?

(iii) Test the null hypothesis that *ros* has no effect on *salary* against the alternative that *ros* has a positive effect. Carry out the test at the 10% significance level.

(iv) Would you include *ros* in a final model explaining CEO compensation in terms of firm performance? Explain.

3 The variable *rdintens* is expenditures on research and development (R&D) as a percentage of sales. Sales are measured in millions of dollars. The variable *profmarg* is profits as a percentage of sales.

Using the data in RDCHEM for 32 firms in the chemical industry, the following equation is estimated:

$$\widehat{rdintens} = .472 + .321 \log(sales) + .050\, profmarg$$
$$(1.369)\ (.216)\qquad (.046)$$
$$n = 32,\ R^2 = .099.$$

(i) Interpret the coefficient on log(*sales*). In particular, if *sales* increases by 10%, what is the estimated percentage point change in *rdintens*? Is this an economically large effect?

(ii) Test the hypothesis that R&D intensity does not change with *sales* against the alternative that it does increase with sales. Do the test at the 5% and 10% levels.

(iii) Interpret the coefficient on *profmarg*. Is it economically large?

(iv) Does *profmarg* have a statistically significant effect on *rdintens* ?

4 Are rent rates influenced by the student population in a college town? Let *rent* be the average monthly rent paid on rental units in a college town in the United States. Let *pop* denote the total city population, *avginc* the average city income, and *pctstu* the student population as a percentage of the total population. One model to test for a relationship is

$$\log(rent) = \beta_0 + \beta_1 \log(pop) + \beta_2 \log(avginc) + \beta_3 pctstu + u.$$

(i) State the null hypothesis that size of the student body relative to the population has no ceteris paribus effect on monthly rents. State the alternative that there is an effect.

(ii) What signs do you expect for β_1 and β_2?

(iii) The equation estimated using 1990 data from RENTAL for 64 college towns is

$$\widehat{\log(rent)} = .043 + .066 \log(pop) + .507 \log(avginc) + .0056 pctstu$$
$$(.844) \quad (.039) \qquad (.081) \qquad (.0017)$$
$$n = 64, R^2 = .458.$$

What is wrong with the statement: "A 10% increase in population is associated with about a 6.6% increase in rent"?

(iv) Test the hypothesis stated in part (i) at the 1% level.

5 Consider the estimated equation from Example 4.3, which can be used to study the effects of skipping class on college GPA:

$$\widehat{colGPA} = 1.39 + .412 hsGPA + .015 ACT - .083 skipped$$
$$(.33) \quad (.094) \qquad (.011) \qquad (.026)$$
$$n = 141, R^2 = .234.$$

(i) Using the standard normal approximation, find the 95% confidence interval for β_{hsGPA}.

(ii) Can you reject the hypothesis H_0: $\beta_{hsGPA} = .4$ against the two-sided alternative at the 5% level?

(iii) Can you reject the hypothesis H_0: $\beta_{hsGPA} = 1$ against the two-sided alternative at the 5% level?

6 In Section 4-5, we used as an example testing the rationality of assessments of housing prices. There, we used a log-log model in *price* and *assess* [see equation (4.47)]. Here, we use a level-level formulation.

(i) In the simple regression model

$$price = \beta_0 + \beta_1 assess + u,$$

the assessment is rational if $\beta_1 = 1$ and $\beta_0 = 0$. The estimated equation is

$$\widehat{price} = -14.47 + .976 assess$$
$$(16.27) \quad (.049)$$
$$n = 88, \text{SSR} = 165{,}644.51, R^2 = .820.$$

First, test the hypothesis that H_0: $\beta_0 = 0$ against the two-sided alternative. Then, test H_0: $\beta_1 = 1$ against the two-sided alternative. What do you conclude?

(ii) To test the joint hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$, we need the SSR in the restricted model. This amounts to computing $\sum_{i=1}^{n}(price_i - assess_i)^2$, where $n = 88$, because the residuals in the restricted model are just $price_i - assess_i$. (No estimation is needed for the restricted model because both parameters are specified under H_0.) This turns out to yield SSR = 209,448.99. Carry out the F test for the joint hypothesis.

(iii) Now, test H_0: $\beta_2 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$ in the model

$$price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u.$$

The R-squared from estimating this model using the same 88 houses is .829.

(iv) If the variance of *price* changes with *assess*, *lotsize*, *sqrft*, or *bdrms*, what can you say about the F test from part (iii)?

7 In Example 4.7, we used data on nonunionized manufacturing firms to estimate the relationship between the scrap rate and other firm characteristics. We now look at this example more closely and use all available firms.

(i) The population model estimated in Example 4.7 can be written as

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales) + \beta_3 \log(employ) + u.$$

Using the 43 observations available for 1987, the estimated equation is

$$\widehat{\log(scrap)} = 11.74 - .042\ hrsemp - .951 \log(sales) + .992 \log(employ)$$
$$(4.57)\ (.019)\qquad (.370)\qquad (.360)$$
$$n = 43,\ R^2 = .310.$$

Compare this equation to that estimated using only the 29 nonunionized firms in the sample.

(ii) Show that the population model can also be written as

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales/employ) + \theta_3 \log(employ) + u,$$

where $\theta_3 = \beta_2 + \beta_3$. [*Hint:* Recall that $\log(x_2/x_3) = \log(x_2) - \log(x_3)$.] Interpret the hypothesis H_0: $\theta_3 = 0$.

(iii) When the equation from part (ii) is estimated, we obtain

$$\widehat{\log(scrap)} = 11.74 - .042\ hrsemp - .951\ \log(sales/employ) + .041\ \log(employ)$$
$$(4.57)\ (.019)\qquad (.370)\qquad (.205)$$
$$n = 43,\ R^2 = .310.$$

Controlling for worker training and for the sales-to-employee ratio, do bigger firms have larger statistically significant scrap rates?

(iv) Test the hypothesis that a 1% increase in *sales*/*employ* is associated with a 1% drop in the scrap rate.

8 Consider the multiple regression model with three independent variables, under the classical linear model assumptions MLR.1 through MLR.6:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

You would like to test the null hypothesis H_0: $\beta_1 - 3\beta_2 = 1$.

(i) Let $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the OLS estimators of β_1 and β_2. Find $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2)$ in terms of the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the covariance between them. What is the standard error of $\hat{\beta}_1 - 3\hat{\beta}_2$?

(ii) Write the t statistic for testing H_0: $\beta_1 - 3\beta_2 = 1$.

(iii) Define $\theta_1 = \beta_1 - 3\beta_2$ and $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$. Write a regression equation involving β_0, θ_1, β_2, and β_3 that allows you to directly obtain $\hat{\theta}_1$ and its standard error.

9 In Problem 3 in Chapter 3, we estimated the equation

$$\widehat{sleep} = 3{,}638.25 - .148\ totwrk - 11.13\ educ + 2.20\ age$$
$$(112.28)\ (.017)\qquad (5.88)\qquad (1.45)$$
$$n = 706,\ R^2 = .113,$$

where we now report standard errors along with the estimates.

(i) Is either *educ* or *age* individually significant at the 5% level against a two-sided alternative? Show your work.

(ii) Dropping *educ* and *age* from the equation gives

$$\widehat{sleep} = 3{,}586.38 - .151\ totwrk$$
$$(38.91)\ \ (.017)$$
$$n = 706,\ R^2 = .103.$$

Are *educ* and *age* jointly significant in the original equation at the 5% level? Justify your answer.

(iii) Does including *educ* and *age* in the model greatly affect the estimated tradeoff between sleeping and working?

(iv) Suppose that the sleep equation contains heteroskedasticity. What does this mean about the tests computed in parts (i) and (ii)?

10 Regression analysis can be used to test whether the market efficiently uses information in valuing stocks. For concreteness, let *return* be the total return from holding a firm's stock over the four-year period from the end of 1990 to the end of 1994. The *efficient markets hypothesis* says that these returns should not be systematically related to information known in 1990. If firm characteristics known at the beginning of the period help to predict stock returns, then we could use this information in choosing stocks.

For 1990, let *dkr* be a firm's debt to capital ratio, let *eps* denote the earnings per share, let *netinc* denote net income, and let *salary* denote total compensation for the CEO.

(i) Using the data in RETURN, the following equation was estimated:

$$\widehat{return} = -14.37 + .321\ dkr + .043\ eps - .0051\ nentinc + .0035\ salary$$
$$(6.89)\ (.201)\ \ (.078)\ \ (.0047)\ \ (.0022)$$
$$n = 142,\ R^2 = .0395.$$

Test whether the explanatory variables are jointly significant at the 5% level. Is any explanatory variable individually significant?

(ii) Now, reestimate the model using the log form for *netinc* and *salary*:

$$\widehat{return} = -36.30 + .327\ dkr + .069\ eps - 4.74\ \log(netinc) + 7.24\ \log(salary)$$
$$(39.37)\ (.203)\ \ (.080)\ \ (3.39)\ \ (6.31)$$
$$n = 142,\ R^2 = .0330.$$

Do any of your conclusions from part (i) change?

(iii) In this sample, some firms have zero debt and others have negative earnings. Should we try to use log(*dkr*) or log(*eps*) in the model to see if these improve the fit? Explain.

(iv) Overall, is the evidence for predictability of stock returns strong or weak?

11 The following table was created using the data in CEOSAL2, where standard errors are in parentheses below the coefficients:

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	——	.112 (.050)	.100 (.049)
profmarg	——	–.0023 (.0022)	–.0022 (.0021)

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
ceoten	——	——	.0171 (.0055)
comten	——	——	–.0092 (.0033)
intercept	4.94 (0.20)	4.62 (0.25)	4.57 (0.25)
Observations	177	177	177
R-squared	.281	.304	.353

The variable *mktval* is market value of the firm, *profmarg* is profit as a percentage of sales, *ceoten* is years as CEO with the current company, and *comten* is total years with the company.

(i) Comment on the effect of *profmarg* on CEO salary.

(ii) Does market value have a significant effect? Explain.

(iii) Interpret the coefficients on *ceoten* and *comten*. Are these explanatory variables statistically significant?

(iv) What do you make of the fact that longer tenure with the company, holding the other factors fixed, is associated with a lower salary?

12 The following analysis was obtained using data in MEAP93, which contains school-level pass rates (as a percent) on a tenth-grade math test.

(i) The variable *expend* is expenditures per student, in dollars, and *math*10 is the pass rate on the exam. The following simple regression relates *math*10 to *lexpend* = log(*expend*):

$$\widehat{math10} = -69.34 + 11.16\ lexpend$$
$$(25.53)\quad(3.17)$$
$$n = 408,\ R^2 = .0297.$$

Interpret the coefficient on *lexpend*. In particular, if *expend* increases by 10%, what is the estimated percentage point change in *math*10? What do you make of the large negative intercept estimate? (The minimum value of *lexpend* is 8.11 and its average value is 8.37.)

(ii) Does the small *R*-squared in part (i) imply that spending is correlated with other factors affecting *math*10? Explain. Would you expect the *R*-squared to be much higher if expenditures were randomly assigned to schools—that is, independent of other school and student characteristics—rather than having the school districts determine spending?

(iii) When log of enrollment and the percent of students eligible for the federal free lunch program are included, the estimated equation becomes

$$\widehat{math10} = -23.14 + 7.75\ lexpend - 1.26\ lenroll - .324\ lnchprg$$
$$(24.99)\quad(3.04)\qquad(0.58)\qquad(0.36)$$
$$n = 408,\ R^2 = .1893.$$

Comment on what happens to the coefficient on *lexpend*. Is the spending coefficient still statistically different from zero?

(iv) What do you make of the *R*-squared in part (iii)? What are some other factors that could be used to explain *math*10 (at the school level)?

13 The data in MEAPSINGLE were used to estimate the following equations relating school-level performance on a fourth-grade math test to socioeconomic characteristics of students attending school. The variable *free*,

measured at the school level, is the percentage of students eligible for the federal free lunch program. The variable *medinc* is median income in the ZIP code, and *pctsgle* is percent of students not living with two parents (also measured at the ZIP code level). See also Computer Exercise C11 in Chapter 3.

$$\widehat{math4} = 96.77 - .833\ pctsgle$$
$$(1.60)\ (.071)$$
$$n = 299,\ R^2 = .380$$

$$\widehat{math4} = 93.00 - .275\ pctsgle - .402\ free$$
$$(1.63)\ (.117)\qquad (.070)$$
$$n = 299,\ R^2 = .459$$

$$\widehat{math4} = 24.49 - .274\ pctsgle - .422\ free - .752\ lmedinc + 9.01\ lexppp$$
$$(59.24)\ (.161)\qquad (.071)\qquad (5.358)\qquad (4.04)$$
$$n = 299,\ R^2 = .472$$

$$\widehat{math4} = 17.52 - .259\ pctsgle - .420\ free + 8.80\ lexppp$$
$$(32.25)\ (.117)\qquad (.070)\qquad (3.76)$$
$$n = 299,\ R^2 = .472.$$

(i) Interpret the coefficient on the variable *pctsgle* in the first equation. Comment on what happens when *free* is added as an explanatory variable.

(ii) Does expenditure per pupil, entered in logarithmic form, have a statistically significant effect on performance? How big is the estimated effect?

(iii) If you had to choose among the four equations as your best estimate of the effect of *pctsgle* and obtain a 95% confidence interval of $\beta_{pctsgle}$, which would you choose? Why?

Computer Exercises

C1 The following model can be used to study whether campaign expenditures affect election outcomes:

$$voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u,$$

where *voteA* is the percentage of the vote received by Candidate A, *expendA* and *expendB* are campaign expenditures by Candidates A and B, and *prtystrA* is a measure of party strength for Candidate A (the percentage of the most recent presidential vote that went to A's party).

(i) What is the interpretation of β_1?

(ii) In terms of the parameters, state the null hypothesis that a 1% increase in A's expenditures is offset by a 1% increase in B's expenditures.

(iii) Estimate the given model using the data in VOTE1 and report the results in usual form. Do A's expenditures affect the outcome? What about B's expenditures? Can you use these results to test the hypothesis in part (ii)?

(iv) Estimate a model that directly gives the t statistic for testing the hypothesis in part (ii). What do you conclude? (Use a two-sided alternative.)

C2 Use the data in LAWSCH85 for this exercise.

(i) Using the same model as in Problem 4 in Chapter 3, state and test the null hypothesis that the rank of law schools has no ceteris paribus effect on median starting salary.

(ii) Are features of the incoming class of students—namely, *LSAT* and *GPA*—individually or jointly significant for explaining *salary*? (Be sure to account for missing data on *LSAT* and *GPA*.)

(iii) Test whether the size of the entering class (*clsize*) or the size of the faculty (*faculty*) needs to be added to this equation; carry out a single test. (Be careful to account for missing data on *clsize* and *faculty*.)

(iv) What factors might influence the rank of the law school that are not included in the salary regression?

C3 Refer to Computer Exercise C2 in Chapter 3. Now, use the log of the housing price as the dependent variable:

$$\log(price) = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u.$$

(i) You are interested in estimating and obtaining a confidence interval for the percentage change in *price* when a 150-square-foot bedroom is added to a house. In decimal form, this is $\theta_1 = 150\beta_1 + \beta_2$. Use the data in HPRICE1 to estimate θ_1.

(ii) Write β_2 in terms of θ_1 and β_1 and plug this into the log(*price*) equation.

(iii) Use part (ii) to obtain a standard error for $\hat{\theta}_1$ and use this standard error to construct a 95% confidence interval.

C4 In Example 4.9, the restricted version of the model can be estimated using all 1,388 observations in the sample. Compute the *R*-squared from the regression of *bwght* on *cigs*, *parity*, and *faminc* using all observations. Compare this to the *R*-squared reported for the restricted model in Example 4.9.

C5 Use the data in MLB1 for this exercise.

(i) Use the model estimated in equation (4.31) and drop the variable *rbisyr*. What happens to the statistical significance of *hrunsyr*? What about the size of the coefficient on *hrunsyr*?

(ii) Add the variables *runsyr* (runs per year), *fldperc* (fielding percentage), and *sbasesyr* (stolen bases per year) to the model from part (i). Which of these factors are individually significant?

(iii) In the model from part (ii), test the joint significance of *bavg*, *fldperc*, and *sbasesyr*.

C6 Use the data in WAGE2 for this exercise.

(i) Consider the standard wage equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

State the null hypothesis that another year of general workforce experience has the same effect on log(*wage*) as another year of tenure with the current employer.

(ii) Test the null hypothesis in part (i) against a two-sided alternative, at the 5% significance level, by constructing a 95% confidence interval. What do you conclude?

C7 Refer to the example used in Section 4-4. You will use the data set TWOYEAR.

(i) The variable *phsrank* is the person's high school percentile. (A higher number is better. For example, 90 means you are ranked better than 90 percent of your graduating class.) Find the smallest, largest, and average *phsrank* in the sample.

(ii) Add *phsrank* to equation (4.26) and report the OLS estimates in the usual form. Is *phsrank* statistically significant? How much is 10 percentage points of high school rank worth in terms of wage?

(iii) Does adding *phsrank* to (4.26) substantively change the conclusions on the returns to two- and four-year colleges? Explain.

(iv) The data set contains a variable called *id*. Explain why if you add *id* to equation (4.17) or (4.26) you expect it to be statistically insignificant. What is the two-sided *p*-value?

C8 The data set 401KSUBS contains information on net financial wealth (*nettfa*), age of the survey respondent (*age*), annual family income (*inc*), family size (*fsize*), and participation in certain pension plans for people in the United States. The wealth and income variables are both recorded in thousands of dollars. For this question, use only the data for single-person households (so *fsize* = 1).

(i) How many single-person households are there in the data set?

(ii) Use OLS to estimate the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u,$$

and report the results using the usual format. Be sure to use only the single-person households in the sample. Interpret the slope coefficients. Are there any surprises in the slope estimates?

(iii) Does the intercept from the regression in part (ii) have an interesting meaning? Explain.

(iv) Find the p-value for the test H_0: $\beta_2 = 1$ against H_1: $\beta_2 < 1$. Do you reject H_0 at the 1% significance level?

(v) If you do a simple regression of *nettfa* on *inc*, is the estimated coefficient on *inc* much different from the estimate in part (ii)? Why or why not?

C9 Use the data in DISCRIM to answer this question. (See also Computer Exercise C8 in Chapter 3.)

(i) Use OLS to estimate the model

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + \beta_3 prppov + u,$$

and report the results in the usual form. Is $\hat{\beta}_1$ statistically different from zero at the 5% level against a two-sided alternative? What about at the 1% level?

(ii) What is the correlation between log(*income*) and *prppov*? Is each variable statistically significant in any case? Report the two-sided p-values.

(iii) To the regression in part (i), add the variable log(*hseval*). Interpret its coefficient and report the two-sided p-value for H_0: $\beta_{\log(hseval)} = 0$.

(iv) In the regression in part (iii), what happens to the individual statistical significance of log(*income*) and *prppov*? Are these variables jointly significant? (Compute a p-value.) What do you make of your answers?

(v) Given the results of the previous regressions, which one would you report as most reliable in determining whether the racial makeup of a zip code influences local fast-food prices?

C10 Use the data in ELEM94_95 to answer this question. The findings can be compared with those in Table 4.1. The dependent variable *lavgsal* is the log of average teacher salary and *bs* is the ratio of average benefits to average salary (by school).

(i) Run the simple regression of *lavgsal* on *bs*. Is the estimated slope statistically different from zero? Is it statistically different from -1?

(ii) Add the variables *lenrol* and *lstaff* to the regression from part (i). What happens to the coefficient on *bs*? How does the situation compare with that in Table 4.1?

(iii) Why is the standard error on the *bs* coefficient smaller in part (ii) than in part (i)? (*Hint:* What happens to the error variance versus multicollinearity when *lenrol* and *lstaff* are added?)

(iv) How come the coefficient on *lstaff* is negative? Is it large in magnitude?

(v) Now add the variable *lunch* to the regression. Holding other factors fixed, are teachers being compensated for teaching students from disadvantaged backgrounds? Explain.

(vi) Overall, is the pattern of results that you find with ELEM94_95 consistent with the pattern in Table 4.1?

C11 Use the data in HTV to answer this question. See also Computer Exercise C10 in Chapter 3.

(i) Estimate the regression model

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u$$

by OLS and report the results in the usual form. Test the null hypothesis that *educ* is linearly related to *abil* against the alternative that the relationship is quadratic.

(ii) Using the equation in part (i), test H_0: $\beta_1 = \beta_2$ against a two-sided alternative. What is the p-value of the test?

(iii) Add the two college tuition variables to the regression from part (i) and determine whether they are jointly statistically significant.

(iv) What is the correlation between *tuit*17 and *tuit*18? Explain why using the average of the tuition over the two years might be preferred to adding each separately. What happens when you do use the average?

(v) Do the findings for the average tuition variable in part (iv) make sense when interpreted causally? What might be going on?

C12 Use the data in ECONMATH to answer the following questions.

(i) Estimate a model explaining *colgpa* to *hsgpa*, *actmth*, and *acteng*. Report the results in the usual form. Are all explanatory variables statistically significant?

(ii) Consider an increase in *hsgpa* of one standard deviation, about .343. By how much does $\widehat{colgpa}$ increase, holding *actmth* and *acteng* fixed. About how many standard deviations would the *actmth* have to increase to change $\widehat{colgpa}$ by the same amount as one standard deviation in *hsgpa*? Comment.

(iii) Test the null hypothesis that *actmth* and *acteng* have the same effect (in the population) against a two-sided alternative. Report the *p*-value and describe your conclusions.

(iv) Suppose the college admissions officer wants you to use the data on the variables in part (i) to create an equation that explains at least 50 percent of the variation in *colgpa*. What would you tell the officer?

C13 Use the data set GPA1 to answer this question. It was used in Computer Exercise C13 in Chapter 3 to estimate the effect of PC ownership on college GPA.

(i) Run the regression *colGPA* on *PC*, *hsGPA*, and *ACT* and obtain a 95% confidence interval for β_{PC}. Is the estimated coefficient statistically significant at the 5% level against a two-sided alternative?

(ii) Discuss the statistical significance of the estimates $\hat{\beta}_{hsGPA}$ and $\hat{\beta}_{ACT}$ in part (i). Is *hsGPA* or *ACT* the more important predictor of *colGPA?*

(iii) Add the two indicators *fathcoll* and *mothcoll* to the regression in part (i). Is either individually significant? Are they jointly statistically significant at the 5% level?

C14 Use the data set JTRAIN98 to answer this question.

(i) Add the unemployment indicator *unem*96 to the regression reported in equation (4.52). Interpret its coefficient and discuss whether its sign and magnitude seem sensible. Is the estimate statistically significant?

(ii) What happens to the estimated job training effect compared with equation (4.52)? Is it still economically and statistically significant?

(iii) Find the correlation between *earn*96 and *unem*96. Is it about what you would expect? Explain.

(iv) Do you think your finding in part (iii) means you should drop *unem*96 from the regression? Explain.

CHAPTER 5

Multiple Regression Analysis: OLS Asymptotics

In Chapters 3 and 4, we covered what are called *finite sample*, *small sample*, or *exact* properties of the OLS estimators in the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u. \qquad [5.1]$$

For example, the unbiasedness of OLS (derived in Chapter 3) under the first four Gauss-Markov assumptions is a finite sample property because it holds for *any* sample size n (subject to the mild restriction that n must be at least as large as the total number of parameters in the regression model, $k + 1$). Similarly, the fact that OLS is the best linear unbiased estimator under the full set of Gauss-Markov assumptions (MLR.1 through MLR.5) is a finite sample property.

In Chapter 4, we added the classical linear model Assumption MLR.6, which states that the error term u is normally distributed and independent of the explanatory variables. This allowed us to derive the *exact* sampling distributions of the OLS estimators (conditional on the explanatory variables in the sample). In particular, Theorem 4.1 showed that the OLS estimators have normal sampling distributions, which led directly to the t and F distributions for t and F statistics. If the error is not normally distributed, the distribution of a t statistic is not exactly t, and an F statistic does not have an exact F distribution for any sample size.

In addition to finite sample properties, it is important to know the **asymptotic properties** or **large sample properties** of estimators and test statistics. These properties are not defined for a particular sample size; rather, they are defined as the sample size grows without bound. Fortunately, under the assumptions we have made, OLS has satisfactory large sample properties. One practically important

finding is that even without the normality assumption (Assumption MLR.6), t and F statistics have *approximately* t and F distributions, at least in large sample sizes. We discuss this in more detail in Section 5-2, after we cover the consistency of OLS in Section 5-1.

Because the material in this chapter is more difficult to understand, and because one can conduct empirical work without a deep understanding of its contents, this chapter may be skipped. However, we will necessarily refer to large sample properties of OLS when we study discrete response variables in Chapter 7, relax the homoskedasticity assumption in Chapter 8, and delve into estimation with time series data in Part 2. Furthermore, virtually all advanced econometric methods derive their justification using large-sample analysis, so readers who will continue into Part 3 should be familiar with the contents of this chapter.

5-1 Consistency

Unbiasedness of estimators, although important, cannot always be obtained. For example, as we discussed in Chapter 3, the standard error of the regression, $\hat{\sigma}$, is not an unbiased estimator for σ, the standard deviation of the error u, in a multiple regression model. Although the OLS estimators are unbiased under MLR.1 through MLR.4, in Chapter 11 we will find that there are time series regressions where the OLS estimators are not unbiased. Further, in Part 3 of the text, we encounter several other estimators that are biased yet useful.

Although not all useful estimators are unbiased, virtually all economists agree that **consistency** is a minimal requirement for an estimator. The Nobel Prize–winning econometrician Clive W. J. Granger once remarked, "If you can't get it right as n goes to infinity, you shouldn't be in this business." The implication is that, if your estimator of a particular population parameter is not consistent, then you are wasting your time.

There are a few different ways to describe consistency. Formal definitions and results are given in Math Refresher C; here, we focus on an intuitive understanding. For concreteness, let $\hat{\beta}_j$ be the OLS estimator of β_j for some j. For each n, $\hat{\beta}_j$ has a probability distribution (representing its possible values in different random samples of size n). Because $\hat{\beta}_j$ is unbiased under Assumptions MLR.1 through MLR.4, this distribution has mean value β_j. If this estimator is consistent, then the distribution of $\hat{\beta}_j$ becomes more and more tightly distributed around β_j as the sample size grows. As n tends to infinity, the distribution of $\hat{\beta}_j$ collapses to the single point β_j. In effect, this means that we can make our estimator arbitrarily close to β_j if we can collect as much data as we want. This convergence is illustrated in Figure 5.1.

Naturally, for any application, we have a fixed sample size, which is a major reason an asymptotic property such as consistency can be difficult to grasp. Consistency involves a thought experiment about what would happen as the sample size gets large (while, at the same time, we obtain numerous random samples for each sample size). If obtaining more and more data does not generally get us closer to the parameter value of interest, then we are using a poor estimation procedure.

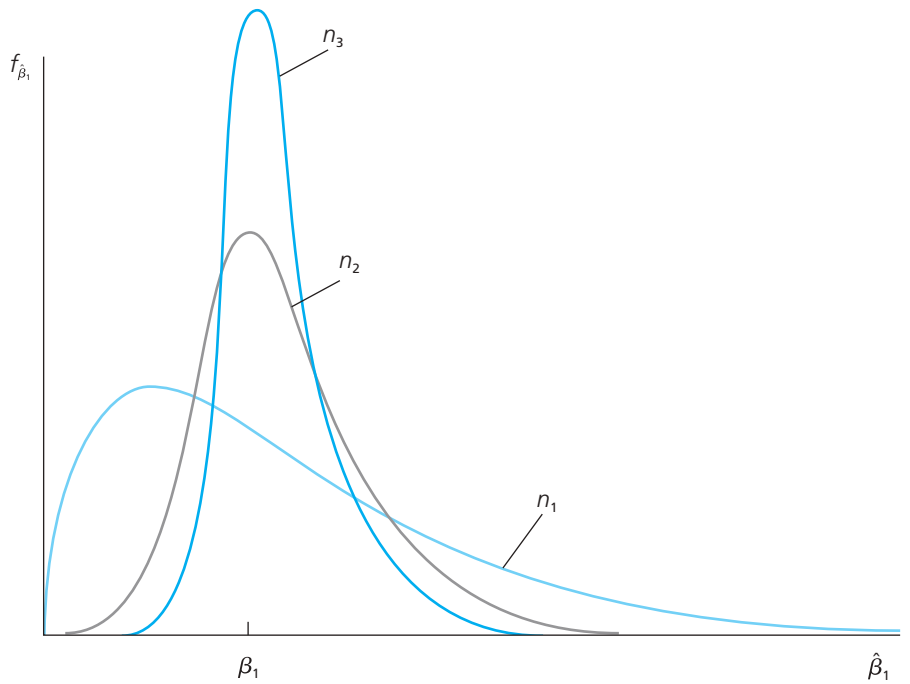
Conveniently, the same set of assumptions implies both unbiasedness and consistency of OLS. We summarize with a theorem.

THEOREM 5.1

CONSISTENCY OF OLS

Under Assumptions MLR.1 through MLR.4, the OLS estimator $\hat{\beta}_j$ is consistent for β_j, for all $j = 0, 1, \ldots, k$.

FIGURE 5.1 Sampling distributions of $\hat{\beta}_1$ for sample sizes $n_1 < n_2 < n_3$.



A general proof of this result is most easily developed using the matrix algebra methods described in Appendices D and E. But we can prove Theorem 5.1 without difficulty in the case of the simple regression model. We focus on the slope estimator, $\hat{\beta}_1$.

The proof starts out the same as the proof of unbiasedness: we write down the formula for $\hat{\beta}_1$, and then plug in $y_i = \beta_0 + \beta_1 x_{i1} + u_i$:

$$\begin{aligned}\hat{\beta}_1 &= \left(\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)y_i\right)\Big/\left(\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2\right) \\ &= \beta_1 + \left(n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)u_i\right)\Big/\left(n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2\right),\end{aligned} \qquad [5.2]$$

where dividing both the numerator and denominator by n does not change the expression but allows us to directly apply the law of large numbers. When we apply the law of large numbers to the averages in the second part of equation (5.2), we conclude that the numerator and denominator converge in probability to the population quantities, $\text{Cov}(x_1,u)$ and $\text{Var}(x_1)$, respectively. Provided that $\text{Var}(x_1) \neq 0$—which is assumed in MLR.3—we can use the properties of *probability limits* (see Math Refresher C) to get

$$\begin{aligned}\text{plim}\,\hat{\beta}_1 &= \beta_1 + \text{Cov}(x_1,u)/\text{Var}(x_1) \\ &= \beta_1 \text{ because } \text{Cov}(x_1,u) = 0.\end{aligned} \qquad [5.3]$$

We have used the fact, discussed in Chapters 2 and 3, that $\text{E}(u|x_1) = 0$ (Assumption MLR.4) implies that x_1 and u are uncorrelated (have zero covariance).

As a technical matter, to ensure that the probability limits exist, we should assume that $\text{Var}(x_1) < \infty$ and $\text{Var}(u) < \infty$ (which means that their probability distributions are not too spread out), but we will not worry about cases where these assumptions might fail. Further, we could—and, in an advanced treatment of econometrics, we would—explicitly relax Assumption MLR.3 to rule out only perfect collinearity in the population. As stated, Assumption MLR.3 also disallows perfect collinearity among the regressors in the sample we have at hand. Technically, for the thought experiment we can show consistency with no perfect collinearity in the population, allowing for the unlucky possibility that we draw a data set that does exhibit perfect collinearity. From a practical perspective the distinction is unimportant, as we cannot compute the OLS estimates for our sample if MLR.3 fails.

The previous arguments, and equation (5.3) in particular, show that OLS is consistent in the simple regression case if we assume only zero correlation. This is also true in the general case. We now state this as an assumption.

Assumption MLR.4′ Zero Mean and Zero Correlation

$\text{E}(u) = 0$ and $\text{Cov}(x_j,u) = 0$, for $j = 1, 2, \ldots, k$.

Assumption MLR.4′ is weaker than Assumption MLR.4 in the sense that the latter implies the former. One way to characterize the zero conditional mean assumption, $\text{E}(u|x_1, \ldots, x_k) = 0$, is that *any* function of the explanatory variables is uncorrelated with u. Assumption MLR.4′ requires only that each x_j is uncorrelated with u (and that u has a zero mean in the population). In Chapter 2, we actually motivated the OLS estimator for simple regression using Assumption MLR.4′, and the first order conditions for OLS in the multiple regression case, given in equation (3.13), are simply the sample analogs of the population zero correlation assumptions (and zero mean assumption). Therefore, in some ways, Assumption MLR.4′ is more natural an assumption because it leads directly to the OLS estimates. Further, when we think about violations of Assumption MLR.4, we usually think in terms of $\text{Cov}(x_j,u) \neq 0$ for some j. So how come we have used Assumption MLR.4 until now? There are two reasons, both of which we have touched on earlier. First, OLS turns out to be biased (but consistent) under Assumption MLR.4′ if $\text{E}(u|x_1, \ldots, x_k)$ depends on any of the x_j. Because we have previously focused on finite sample, or exact, sampling properties of the OLS estimators, we have needed the stronger zero conditional mean assumption.

Second, and probably more important, is that the zero conditional mean assumption means that we have properly modeled the population regression function (PRF). That is, under Assumption MLR.4 we can write

$$\text{E}(y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

and so we can obtain partial effects of the explanatory variables on the average or expected value of y. If we instead only assume Assumption MLR.4′, $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ need not represent the PRF, and we face the possibility that some nonlinear functions of the x_j, such as x_j^2, could be correlated with the error u. A situation like this means that we have neglected nonlinearities in the model that could help us better explain y; if we knew that, we would usually include such nonlinear functions. In other words, most of the time we hope to get a good estimate of the PRF, and so the zero conditional mean assumption is natural. Nevertheless, the weaker zero correlation assumption turns out to be useful in interpreting OLS estimation of a linear model as providing the best linear approximation to the PRF. It is also used in more advanced settings, such as in Chapter 15, where we have no interest in modeling a PRF. For further discussion of this somewhat subtle point, see Wooldridge (2010, Chapter 4) as well as Problem 6 at the end of this chapter.

5-1a Deriving the Inconsistency in OLS

Just as failure of $E(u|x_1, \ldots, x_k) = 0$ causes bias in the OLS estimators, correlation between u and *any* of $x_1, x_2, \ldots, x_k$ generally causes *all* of the OLS estimators to be inconsistent. This simple but important observation is often summarized as: *if the error is correlated with any of the independent variables, then OLS is biased and inconsistent*. This is very unfortunate because it means that any bias persists as the sample size grows.

In the simple regression case, we can obtain the inconsistency from the first part of equation (5.3), which holds whether or not u and x_1 are uncorrelated. The **inconsistency** in $\hat{\beta}_1$ (sometimes loosely called the **asymptotic bias**) is

$$\text{plim}\,\hat{\beta}_1 - \beta_1 = \text{Cov}(x_1,u)/\text{Var}(x_1). \qquad [5.4]$$

Because $\text{Var}(x_1) > 0$, the inconsistency in $\hat{\beta}_1$ is positive if x_1 and u are positively correlated, and the inconsistency is negative if x_1 and u are negatively correlated. If the covariance between x_1 and u is small relative to the variance in x_1, the inconsistency can be negligible; unfortunately, we cannot even estimate how big the covariance is because u is unobserved.

We can use (5.4) to derive the asymptotic analog of the omitted variable bias (see Table 3.2 in Chapter 3). Suppose the true model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v,$$

satisfies the first four Gauss-Markov assumptions. Then v has a zero mean and is uncorrelated with x_1 and x_2. If $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ denote the OLS estimators from the regression of y on x_1 and x_2, then Theorem 5.1 implies that these estimators are consistent. If we omit x_2 from the regression and do the simple regression of y on x_1, then $u = \beta_2 x_2 + v$. Let $\tilde{\beta}_1$ denote the simple regression slope estimator. Then

$$\text{plim}\,\tilde{\beta}_1 = \beta_1 + \beta_2 \delta_1, \qquad [5.5]$$

where

$$\delta_1 = \text{Cov}(x_1,x_2)/\text{Var}(x_1). \qquad [5.6]$$

Thus, for practical purposes, we can view the inconsistency as being the same as the bias. The difference is that the inconsistency is expressed in terms of the population variance of x_1 and the population covariance between x_1 and x_2, while the bias is based on their sample counterparts (because we condition on the values of x_1 and x_2 in the sample).

If x_1 and x_2 are uncorrelated (in the population), then $\delta_1 = 0$, and $\tilde{\beta}_1$ is a consistent estimator of β_1 (although not necessarily unbiased). If x_2 has a positive partial effect on y, so that $\beta_2 > 0$, and x_1 and x_2 are positively correlated, so that $\delta_1 > 0$, then the inconsistency in $\tilde{\beta}_1$ is positive, and so on. We can obtain the direction of the inconsistency or asymptotic bias from Table 3.2. If the covariance between x_1 and x_2 is small relative to the variance of x_1, the inconsistency can be small.

EXAMPLE 5.1 **Housing Prices and Distance from an Incinerator**

Let y denote the price of a house (*price*), let x_1 denote the distance from the house to a new trash incinerator (*distance*), and let x_2 denote the "quality" of the house (*quality*). The variable *quality* is left vague so that it can include things like size of the house and lot, number of bedrooms and bathrooms, and intangibles such as attractiveness of the neighborhood. If the incinerator depresses house prices, then β_1 should be positive: everything else being equal, a house that is farther away from the incinerator is worth more. By definition, β_2 is positive because higher quality houses sell for more, other factors being equal. If the incinerator was built farther away, on average, from better homes, then *distance* and *quality* are positively correlated, and so $\delta_1 > 0$. A simple regression of *price* on *distance* [or log(*price*) on log(*distance*)] will tend to overestimate the effect of the incinerator: $\beta_1 + \beta_2 \delta_1 > \beta_1$.

GOING FURTHER 5.1

Suppose that the model

$$score = \beta_0 + \beta_1 skipped + \beta_2 priGPA + u$$

satisfies the first four Gauss-Markov assumptions, where *score* is score on a final exam, *skipped* is number of classes skipped, and *priGPA* is GPA prior to the current semester. If $\tilde{\beta}_1$ is from the simple regression of *score* on *skipped*, what is the direction of the asymptotic bias in $\tilde{\beta}_1$?

An important point about inconsistency in OLS estimators is that, by definition, the problem does not go away by adding more observations to the sample. If anything, the problem gets worse with more data: the OLS estimator gets closer and closer to $\beta_1 + \beta_2\delta_1$ as the sample size grows.

Deriving the sign and magnitude of the inconsistency in the general k regressor case is harder, just as deriving the bias is more difficult. We need to remember that if we have the model in equation (5.1) where, say, x_1 is correlated with u but the other independent variables are uncorrelated with u, *all* of the OLS estimators are generally inconsistent. For example, in the $k = 2$ case,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

suppose that x_2 and u are uncorrelated but x_1 and u are correlated. Then the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ will generally both be inconsistent. (The intercept will also be inconsistent.) The inconsistency in $\hat{\beta}_2$ arises when x_1 and x_2 are correlated, as is usually the case. If x_1 and x_2 are uncorrelated, then any correlation between x_1 and u does *not* result in the inconsistency of $\hat{\beta}_2$: plim $\hat{\beta}_2 = \beta_2$. Further, the inconsistency in $\hat{\beta}_1$ is the same as in (5.4). The same statement holds in the general case: if x_1 is correlated with u, but x_1 and u are uncorrelated with the other independent variables, then only $\hat{\beta}_1$ is inconsistent, and the inconsistency is given by (5.4). The general case is very similar to the omitted variable case in Section 3A.4 of Appendix 3A.

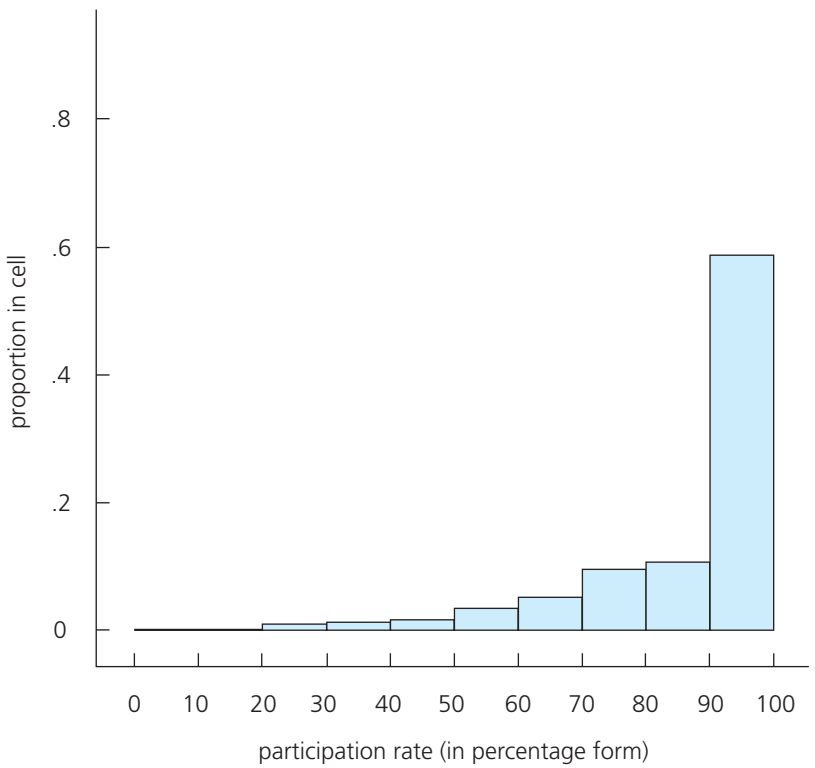
5-2 Asymptotic Normality and Large Sample Inference

Consistency of an estimator is an important property, but it alone does not allow us to perform statistical inference. Simply knowing that the estimator is getting closer to the population value as the sample size grows does not allow us to test hypotheses about the parameters. For testing, we need the sampling distribution of the OLS estimators. Under the classical linear model assumptions MLR.1 through MLR.6, Theorem 4.1 shows that the sampling distributions are normal. This result is the basis for deriving the t and F distributions that we use so often in applied econometrics.

The exact normality of the OLS estimators hinges crucially on the normality of the distribution of the error, u, in the population. If the errors $u_1, u_2, \ldots, u_n$ are random draws from some distribution other than the normal, the $\hat{\beta}_j$ will not be normally distributed, which means that the t statistics will not have t distributions and the F statistics will not have F distributions. This is a potentially serious problem because our inference hinges on being able to obtain critical values or p-values from the t or F distributions.

Recall that Assumption MLR.6 is equivalent to saying that the distribution of y given $x_1, x_2, \ldots, x_k$ is normal. Because y is observed and u is not, in a particular application, it is much easier to think about whether the distribution of y is likely to be normal. In fact, we have already seen a few examples where y definitely cannot have a conditional normal distribution. A normally distributed random variable is symmetrically distributed about its mean, it can take on any positive or negative value, and more than 95% of the area under the distribution is within two standard deviations.

FIGURE 5.2 Histogram of *prate* using the data in 401K.



In Example 3.5, we estimated a model explaining the number of arrests of young men during a particular year (*narr86*). In the population, most men are not arrested during the year, and the vast majority are arrested one time at the most. (In the sample of 2,725 men in the data set CRIME1, fewer than 8% were arrested more than once during 1986.) Because *narr86* takes on only two values for 92% of the sample, it cannot be close to being normally distributed in the population.

In Example 4.6, we estimated a model explaining participation percentages (*prate*) in 401(k) pension plans. The frequency distribution (also called a *histogram*) in Figure 5.2 shows that the distribution of *prate* is heavily skewed to the right, rather than being normally distributed. In fact, over 40% of the observations on *prate* are at the value 100, indicating 100% participation. This violates the normality assumption even conditional on the explanatory variables.

We know that normality plays no role in the unbiasedness of OLS, nor does it affect the conclusion that OLS is the best linear unbiased estimator under the Gauss-Markov assumptions. But exact inference based on t and F statistics requires MLR.6. Does this mean that, in our prior analysis of *prate* in Example 4.6, we must abandon the t statistics for determining which variables are statistically significant? Fortunately, the answer to this question is *no*. Even though the y_i are not from a normal distribution, we can use the central limit theorem from Math Refresher C to conclude that the OLS estimators satisfy **asymptotic normality**, which means they are approximately normally distributed in large enough sample sizes.

THEOREM 5.2

ASYMPTOTIC NORMALITY OF OLS

Under the Gauss-Markov Assumptions MLR.1 through MLR.5,

(i) $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim} \text{Normal}(0, \sigma^2/a_j^2)$, where $\sigma^2/a_j^2 > 0$ is the **asymptotic variance** of $\sqrt{n}(\hat{\beta}_j - \beta_j)$; for the slope coefficients, $a_j^2 = \text{plim}(n^{-1}\sum_{i=1}^{n}\hat{r}_{ij}^2)$, where the $\hat{r}_{ij}$ are the residuals from regressing x_j on the other independent variables. We say that $\hat{\beta}_j$ is *asymptotically normally distributed* (see Math Refresher C);

(ii) $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2 = \text{Var}(u)$;

(iii) For each j,

$$(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \overset{a}{\sim} \text{Normal}(0,1)$$

and

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \overset{a}{\sim} \text{Normal}(0,1), \qquad [5.7]$$

where $\text{se}(\hat{\beta}_j)$ is the usual OLS standard error.

The proof of asymptotic normality is somewhat complicated and is sketched in the appendix for the simple regression case. Part (ii) follows from the law of large numbers, and part (iii) follows from parts (i) and (ii) and the asymptotic properties discussed in Math Refresher C.

Theorem 5.2 is useful because the normality Assumption MLR.6 has been dropped; the only restriction on the distribution of the error is that it has finite variance, something we will always assume. We have also assumed zero conditional mean (MLR.4) and homoskedasticity of u (MLR.5).

In trying to understand the meaning of Theorem 5.2, it is important to keep separate the notions of the population distribution of the error term, u, and the sampling distributions of the $\hat{\beta}_j$ as the sample size grows. A common mistake is to think that something is happening to the distribution of u—namely, that it is getting "closer" to normal—as the sample size grows. But remember that the population distribution is immutable and has nothing to do with the sample size. For example, we previously discussed *narr86*, the number of times a young man is arrested during the year 1986. The nature of this variable—it takes on small, nonnegative integer values—is fixed in the population. Whether we sample 10 men or 1,000 men from this population obviously has no effect on the population distribution.

What Theorem 5.2 says is that, regardless of the population distribution of u, the OLS estimators, when properly standardized, have approximate standard normal distributions. This approximation comes about by the central limit theorem because the OLS estimators involve—in a complicated way—the use of sample averages. Effectively, the sequence of distributions of averages of the underlying errors is approaching normality for virtually any population distribution.

Notice how the standardized $\hat{\beta}_j$ has an asymptotic standard normal distribution whether we divide the difference $\hat{\beta}_j - \beta_j$ by $\text{sd}(\hat{\beta}_j)$ (which we do not observe because it depends on σ) or by $\text{se}(\hat{\beta}_j)$ (which we can compute from our data because it depends on $\hat{\sigma}$). In other words, from an asymptotic point of view it does not matter that we have to replace σ with $\hat{\sigma}$. Of course, replacing σ with $\hat{\sigma}$ affects the exact distribution of the standardized $\hat{\beta}_j$. We just saw in Chapter 4 that under the classical linear model assumptions, $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j)$ has an exact Normal(0,1) distribution and $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ has an exact t_{n-k-1} distribution.

How should we use the result in equation (5.7)? It may seem one consequence is that, if we are going to appeal to large-sample analysis, we should now use the standard normal distribution for inference rather than the t distribution. But from a practical perspective it is just as legitimate to write

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \overset{a}{\sim} t_{n-k-1} = t_{df}, \qquad [5.8]$$

because t_{df} approaches the Normal(0,1) distribution as df gets large. Because we know under the CLM assumptions the t_{n-k-1} distribution holds exactly, it makes sense to treat $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ as a t_{n-k-1} random variable generally, even when MLR.6 does not hold.

Equation (5.8) tells us that t testing and the construction of confidence intervals are carried out *exactly* as under the classical linear model assumptions. This means that our analysis of dependent variables like *prate* and *narr86* does not have to change at all if the Gauss-Markov assumptions hold: in both cases, we have at least 1,500 observations, which is certainly enough to justify the approximation of the central limit theorem.

If the sample size is not very large, then the t distribution can be a poor approximation to the distribution of the t statistics when u is not normally distributed. Unfortunately, there are no general prescriptions on how big the sample size must be before the approximation is good enough. Some econometricians think that $n = 30$ is satisfactory, but this cannot be sufficient for all possible distributions of u. Depending on the distribution of u, more observations may be necessary before the central limit theorem delivers a useful approximation. Further, the quality of the approximation depends not just on n, but on the df, $n - k - 1$: with more independent variables in the model, a larger sample size is usually needed to use the t approximation. Methods for inference with small degrees of freedom and nonnormal errors are outside the scope of this text. We will simply use the t statistics as we always have without worrying about the normality assumption.

It is very important to see that Theorem 5.2 *does* require the homoskedasticity assumption (along with the zero conditional mean assumption). If $\text{Var}(y|\mathbf{x})$ is not constant, the usual t statistics and confidence intervals are invalid no matter how large the sample size is; the central limit theorem does not bail us out when it comes to heteroskedasticity. For this reason, we devote all of Chapter 8 to discussing what can be done in the presence of heteroskedasticity.

One conclusion of Theorem 5.2 is that $\hat{\sigma}^2$ is a consistent estimator of σ^2; we already know from Theorem 3.3 that $\hat{\sigma}^2$ is unbiased for σ^2 under the Gauss-Markov assumptions. The consistency implies that $\hat{\sigma}$ is a consistent estimator of σ, which is important in establishing the asymptotic normality result in equation (5.7).

Remember that $\hat{\sigma}$ appears in the standard error for each $\hat{\beta}_j$. In fact, the estimated variance of $\hat{\beta}_j$ is

$$\widehat{\text{Var}}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\text{SST}_j(1 - R_j^2)}, \qquad [5.9]$$

where SST_j is the total sum of squares of x_j in the sample, and R_j^2 is the R-squared from regressing x_j on all of the other independent variables. In Section 3-4, we studied each component of (5.9), which we will now expound on in the context of asymptotic analysis. As the sample size grows, $\hat{\sigma}^2$ converges in probability to the constant σ^2. Further, R_j^2 approaches a number strictly between zero and unity (so that $1 - R_j^2$ converges to some number between zero and one). The sample variance of x_j is SST_j/n, and so SST_j/n converges to $\text{Var}(x_j)$ as the sample size grows. This means that SST_j grows at approximately the same rate as the sample size: $\text{SST}_j \approx n\sigma_j^2$, where σ_j^2 is the population variance of x_j. When we combine these facts, we find that $\widehat{\text{Var}}(\hat{\beta}_j)$ shrinks to zero at the rate of $1/n$; this is why larger sample sizes are better.

When u is not normally distributed, the square root of (5.9) is sometimes called the **asymptotic standard error**, and t statistics are called **asymptotic *t* statistics**. Because these are the same quantities we dealt with in Chapter 4, we will just call them standard errors and t statistics, with the understanding that sometimes they have only large-sample justification. A similar comment holds for an asymptotic confidence interval constructed from the asymptotic standard error.

GOING FURTHER 5.2

In a regression model with a large sample size, what is an approximate 95% confidence interval for $\hat{\beta}_j$ under MLR.1 through MLR.5? We call this an **asymptotic confidence interval**.

Using the preceding argument about the estimated variance, we can write

$$\text{se}(\hat{\beta}_j) \approx c_j/\sqrt{n}, \qquad [5.10]$$

where c_j is a positive constant that does *not* depend on the sample size. In fact, the constant c_j can be shown to be

$$c_j = \frac{\sigma}{\sigma_j\sqrt{1 - \rho_j^2}},$$

where $\sigma = \text{sd}(u)$, $\sigma_j = \text{sd}(x_j)$, and ρ_j^2 is the population R-squared from regressing x_j on the other explanatory variables. Just like studying equation (5.9) to see which variables affect $\text{Var}(\hat{\beta}_j)$ under the Gauss-Markov assumptions, we can use this expression for c_j to study the impact of larger error standard deviation (σ), more population variation in x_j (σ_j), and multicollinearity in the population (ρ_j^2).

Equation (5.10) is only an approximation, but it is a useful rule of thumb: standard errors can be expected to shrink at a rate that is the inverse of the *square root* of the sample size.

EXAMPLE 5.2 **Standard Errors in a Birth Weight Equation**

We use the data in BWGHT to estimate a relationship where log of birth weight is the dependent variable, and cigarettes smoked per day (*cigs*) and log of family income are independent variables. The total number of observations is 1,388. Using the first half of the observations (694), the standard error for $\hat{\beta}_{cigs}$ is about .0013. The standard error using all of the observations is about .00086. The ratio of the latter standard error to the former is .00086/.0013 ≈ .662. This is pretty close to $\sqrt{694/1{,}388} \approx .707$, the ratio obtained from the approximation in (5.10). In other words, equation (5.10) implies that the standard error using the larger sample size should be about 70.7% of the standard error using the smaller sample. This percentage is pretty close to the 66.2% we actually compute from the ratio of the standard errors.

The asymptotic normality of the OLS estimators also implies that the F statistics have approximate F distributions in large sample sizes. Thus, for testing exclusion restrictions or other multiple hypotheses, nothing changes from what we have done before.

5-2a Other Large Sample Tests: The Lagrange Multiplier Statistic

Once we enter the realm of asymptotic analysis, other test statistics can be used for hypothesis testing. For most purposes, there is little reason to go beyond the usual t and F statistics: as we just saw, these statistics have large sample justification without the normality assumption. Nevertheless, sometimes it is useful to have other ways to test multiple exclusion restrictions, and we now cover the **Lagrange multiplier (*LM*) statistic**, which has achieved some popularity in modern econometrics.

The name "Lagrange multiplier statistic" comes from constrained optimization, a topic beyond the scope of this text. [See Davidson and MacKinnon (1993).] The name **score statistic**—which also comes from optimization using calculus—is used as well. Fortunately, in the linear regression framework, it is simple to motivate the *LM* statistic without delving into complicated mathematics.

The form of the *LM* statistic we derive here relies on the Gauss-Markov assumptions, the same assumptions that justify the F statistic in large samples. We do not need the normality assumption.

To derive the *LM* statistic, consider the usual multiple regression model with k independent variables:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u. \qquad [5.11]$$

We would like to test whether, say, the last q of these variables all have zero population parameters: the null hypothesis is

$$H_0: \beta_{k-q+1} = 0, \ldots, \beta_k = 0, \quad [5.12]$$

which puts q exclusion restrictions on the model (5.11). As with F testing, the alternative to (5.12) is that at least one of the parameters is different from zero.

The *LM* statistic requires estimation of the *restricted* model only. Thus, assume that we have run the regression

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_{k-q} x_{k-q} + \tilde{u}, \quad [5.13]$$

where "~" indicates that the estimates are from the restricted model. In particular, $\tilde{u}$ indicates the residuals from the restricted model. (As always, this is just shorthand to indicate that we obtain the restricted residual for each observation in the sample.)

If the omitted variables x_{k-q+1} through x_k truly have zero population coefficients, then, at least approximately, $\tilde{u}$ should be uncorrelated with each of these variables in the sample. This suggests running a regression of these residuals on those independent variables excluded under H_0, which is almost what the *LM* test does. However, it turns out that, to get a usable test statistic, we must include *all* of the independent variables in the regression. (We must include all regressors because, in general, the omitted regressors in the restricted model are correlated with the regressors that appear in the restricted model.) Thus, we run the regression of

$$\tilde{u} \text{ on } x_1, x_2, \ldots, x_k. \quad [5.14]$$

This is an example of an **auxiliary regression**, a regression that is used to compute a test statistic but whose coefficients are not of direct interest.

How can we use the regression output from (5.14) to test (5.12)? If (5.12) is true, the R-squared from (5.14) should be "close" to zero, subject to sampling error, because $\tilde{u}$ will be approximately uncorrelated with all the independent variables. The question, as always with hypothesis testing, is how to determine when the statistic is large enough to reject the null hypothesis at a chosen significance level. It turns out that, under the null hypothesis, the sample size multiplied by the usual R-squared from the auxiliary regression (5.14) is distributed asymptotically as a chi-square random variable with q degrees of freedom. This leads to a simple procedure for testing the joint significance of a set of q independent variables.

The Lagrange Multiplier Statistic for q Exclusion Restrictions:

(i) Regress y on the *restricted* set of independent variables and save the residuals, $\tilde{u}$.

(ii) Regress $\tilde{u}$ on *all* of the independent variables and obtain the R-squared, say, R_u^2 (to distinguish it from the R-squareds obtained with y as the dependent variable).

(iii) Compute $LM = nR_u^2$ [the sample size times the R-squared obtained from step (ii)].

(iv) Compare LM to the appropriate critical value, c, in a χ_q^2 distribution; if $LM > c$, the null hypothesis is rejected. Even better, obtain the p-value as the probability that a χ_q^2 random variable exceeds the value of the test statistic. If the p-value is less than the desired significance level, then H_0 is rejected. If not, we fail to reject H_0. The rejection rule is essentially the same as for F testing.

Because of its form, the *LM* statistic is sometimes referred to as the ***n*-*R*-squared statistic**. Unlike with the F statistic, the degrees of freedom in the unrestricted model plays no role in carrying out the *LM* test. All that matters is the number of restrictions being tested (q), the size of the auxiliary R-squared (R_u^2), and the sample size (n). The df in the unrestricted model plays no role because of the

asymptotic nature of the *LM* statistic. But we must be sure to multiply R_u^2 by the sample size to obtain *LM*; a seemingly low value of the *R*-squared can still lead to joint significance if *n* is large.

Before giving an example, a word of caution is in order. If in step (i), we mistakenly regress *y* on all of the independent variables and obtain the residuals from this unrestricted regression to be used in step (ii), we do not get an interesting statistic: the resulting *R*-squared will be exactly zero! This is because OLS chooses the estimates so that the residuals are uncorrelated in samples with all included independent variables [see equations in (3.13)]. Thus, we can only test (5.12) by regressing the restricted residuals on *all* of the independent variables. (Regressing the restricted residuals on the restricted set of independent variables will also produce $R^2 = 0$.)

EXAMPLE 5.3 Economic Model of Crime

We illustrate the *LM* test by using a slight extension of the crime model from Example 3.5:

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

where

$narr86$ = the number of times a man was arrested.
$pcnv$ = the proportion of prior arrests leading to conviction.
$avgsen$ = average sentence served from past convictions.
$tottime$ = total time the man has spent in prison prior to 1986 since reaching the age of 18.
$ptime86$ = months spent in prison in 1986.
$qemp86$ = number of quarters in 1986 during which the man was legally employed.

We use the *LM* statistic to test the null hypothesis that *avgsen* and *tottime* have no effect on *narr86* once the other factors have been controlled for.

In step (i), we estimate the restricted model by regressing *narr86* on *pcnv*, *ptime86*, and *qemp86*; the variables *avgsen* and *tottime* are excluded from this regression. We obtain the residuals $\tilde{u}$ from this regression, 2,725 of them. Next, we run the regression of

$$\tilde{u} \text{ on } pcnv, ptime86, qemp86, avgsen, \text{ and } tottime; \qquad [5.15]$$

as always, the order in which we list the independent variables is irrelevant. This second regression produces R_u^2, which turns out to be about .0015. This may seem small, but we must multiply it by *n* to get the *LM* statistic: $LM = 2{,}725(.0015) \approx 4.09$. The 10% critical value in a chi-square distribution with two degrees of freedom is about 4.61 (rounded to two decimal places; see Table G.4). Thus, we fail to reject the null hypothesis that $\beta_{avgsen} = 0$ and $\beta_{tottime} = 0$ at the 10% level. The *p*-value is $P(\chi_2^2 > 4.09) \approx .129$, so we would reject H_0 at the 15% level.

As a comparison, the *F* test for joint significance of *avgsen* and *tottime* yields a *p*-value of about .131, which is pretty close to that obtained using the *LM* statistic. This is not surprising because, asymptotically, the two statistics have the same probability of Type I error. (That is, they reject the null hypothesis with the same frequency when the null is true.)

As the previous example suggests, with a large sample, we rarely see important discrepancies between the outcomes of *LM* and *F* tests. We will use the *F* statistic for the most part because it is computed routinely by most regression packages. But you should be aware of the *LM* statistic as it is used in applied work.

One final comment on the *LM* statistic. As with the *F* statistic, we must be sure to use the same observations in steps (i) and (ii). If data are missing for some of the independent variables that are excluded under the null hypothesis, the residuals from step (i) should be obtained from a regression on the reduced data set.

5-3 Asymptotic Efficiency of OLS

We know that, under the Gauss-Markov assumptions, the OLS estimators are best linear unbiased. OLS is also **asymptotically efficient** among a certain class of estimators under the Gauss-Markov assumptions. A general treatment requires matrix algebra and advanced asymptotic analysis. First, we describe the result in the simple regression case.

In the model

$$y = \beta_0 + \beta_1 x + u, \tag{5.16}$$

u has a zero conditional mean under MLR.4: $\mathrm{E}(u|x) = 0$. This opens up a variety of consistent estimators for β_0 and β_1; as usual, we focus on the slope parameter, β_1. Let $g(x)$ be any function of x; for example, $g(x) = x^2$ or $g(x) = 1/(1 + |x|)$. Then u is uncorrelated with $g(x)$ (see Property CE.5 in Math Refresher B). Let $z_i = g(x_i)$ for all observations i. Then the estimator

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n}(z_i - \bar{z})y_i\right)\Big/\left(\sum_{i=1}^{n}(z_i - \bar{z})x_i\right) \tag{5.17}$$

is consistent for β_1, provided $g(x)$ and x are correlated. [Remember, it is possible that $g(x)$ and x are uncorrelated because correlation measures *linear* dependence.] To see this, we can plug in $y_i = \beta_0 + \beta_1 x_i + u_i$ and write $\tilde{\beta}_1$ as

$$\tilde{\beta}_1 = \beta_1 + \left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})u_i\right)\Big/\left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})x_i\right). \tag{5.18}$$

Now, we can apply the law of large numbers to the numerator and denominator, which converge in probability to $\mathrm{Cov}(z,u)$ and $\mathrm{Cov}(z,x)$, respectively. Provided that $\mathrm{Cov}(z,u) \neq 0$—so that z and x are correlated—we have

$$\operatorname{plim} \tilde{\beta}_1 = \beta_1 + \mathrm{Cov}(z,u)/\mathrm{Cov}(z,x) = \beta_1,$$

because $\mathrm{Cov}(z,u) = 0$ under MLR.4.

It is more difficult to show that $\tilde{\beta}_1$ is asymptotically normal. Nevertheless, using arguments similar to those in the appendix, it can be shown that $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$ is asymptotically normal with mean zero and asymptotic variance $\sigma^2\mathrm{Var}(z)/[\mathrm{Cov}(z,x)]^2$. The asymptotic variance of the OLS estimator is obtained when $z = x$, in which case, $\mathrm{Cov}(z,x) = \mathrm{Cov}(x,x) = \mathrm{Var}(x)$. Therefore, the asymptotic variance of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$, where $\hat{\beta}_1$ is the OLS estimator, is $\sigma^2\mathrm{Var}(x)/[\mathrm{Var}(x)]^2 = \sigma^2/\mathrm{Var}(x)$.Now, the Cauchy-Schwartz inequality (see Math Refresher B.4) implies that $[\mathrm{Cov}(z,x)]^2 \leq \mathrm{Var}(z)\mathrm{Var}(x)$, which implies that the asymptotic variance of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is no larger than that of $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$. We have shown in the simple regression case that, under the Gauss-Markov assumptions, the OLS estimator has a smaller asymptotic variance than any estimator of the form (5.17). [The estimator in (5.17) is an example of an *instrumental variables estimator*, which we will study extensively in Chapter 15.] If the homoskedasticity assumption fails, then there are estimators of the form (5.17) that have a smaller asymptotic variance than OLS. We will see this in Chapter 8.

The general case is similar but much more difficult mathematically. In the k regressor case, the class of consistent estimators is obtained by generalizing the OLS first order conditions:

$$\sum_{i=1}^{n} g_j(\boldsymbol{x}_i)(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_k x_{ik}) = 0, j = 0, 1, \ldots, k, \tag{5.19}$$

where $g_j(\boldsymbol{x}_i)$ denotes any function of all explanatory variables for observation i. As can be seen by comparing (5.19) with the OLS first order conditions in (3.13), we obtain the OLS estimators when $g_0(\boldsymbol{x}_i) = 1$ and $g_j(\boldsymbol{x}_i) = x_{ij}$ for $j = 1, 2, \ldots, k$. There are infinitely many estimators that can be defined using the equations in (5.19) because we can use any functions of the x_{ij} that we want.

THEOREM 5.3 **ASYMPTOTIC EFFICIENCY OF OLS**

Under the Gauss-Markov assumptions, let $\tilde{\beta}_j$ denote estimators that solve equations of the form (5.19) and let $\hat{\beta}_j$ denote the OLS estimators. Then for $j = 0, 1, 2, \ldots, k$, the OLS estimators have the smallest asymptotic variances: $\text{Avar}\sqrt{n}(\hat{\beta}_j - \beta_j) \leq \text{Avar}\sqrt{n}(\tilde{\beta}_j - \beta_j)$.

Proving consistency of the estimators in (5.19), let alone showing they are asymptotically normal, is mathematically difficult. See Wooldridge (2010, Chapter 5).

Summary

The claims underlying the material in this chapter are fairly technical, but their practical implications are straightforward. We have shown that the first four Gauss-Markov assumptions imply that OLS is consistent. Furthermore, all of the methods of testing and constructing confidence intervals that we learned in Chapter 4 are approximately valid without assuming that the errors are drawn from a normal distribution (equivalently, the distribution of y given the explanatory variables is not normal). This means that we can apply OLS and use previous methods for an array of applications where the dependent variable is not even approximately normally distributed. We also showed that the *LM* statistic can be used instead of the *F* statistic for testing exclusion restrictions.

Before leaving this chapter, we should note that examples such as Example 5.3 may very well have problems that *do* require special attention. For a variable such as *narr86*, which is zero or one for most men in the population, a linear model may not be able to adequately capture the functional relationship between *narr86* and the explanatory variables. Moreover, even if a linear model does describe the expected value of arrests, heteroskedasticity might be a problem. Problems such as these are not mitigated as the sample size grows, and we will return to them in later chapters.

Key Terms

Asymptotic Bias
Asymptotic Confidence Interval
Asymptotic Normality
Asymptotic Properties
Asymptotic Standard Error
Asymptotic *t* Statistics
Asymptotic Variance
Asymptotically Efficient
Auxiliary Regression
Consistency
Inconsistency
Lagrange Multiplier (*LM*) Statistic
Large Sample Properties
n-*R*-Squared Statistic
Score Statistic

Problems

1 In the simple regression model under MLR.1 through MLR.4, we argued that the slope estimator, $\hat{\beta}_1$, is consistent for β_1. Using $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1$, show that plim $\hat{\beta}_0 = \beta_0$. [You need to use the consistency of $\hat{\beta}_1$ and the law of large numbers, along with the fact that $\beta_0 = \text{E}(y) - \beta_1\text{E}(x_1)$.]

2 Suppose that the model

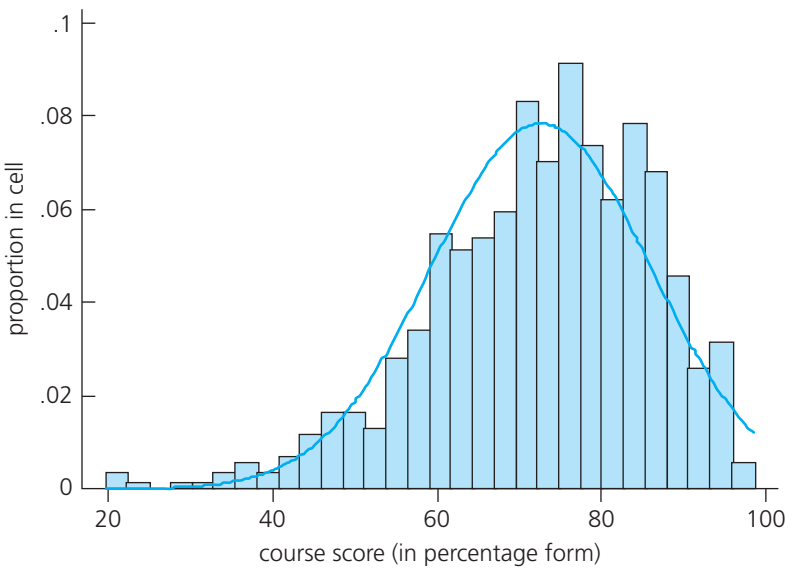
$$pctstck = \beta_0 + \beta_1 funds + \beta_2 risktol + u$$

satisfies the first four Gauss-Markov assumptions, where *pctstck* is the percentage of a worker's pension invested in the stock market, *funds* is the number of mutual funds that the worker can choose from, and *risktol* is some measure of risk tolerance (larger *risktol* means the person has a higher tolerance for risk). If *funds* and *risktol* are positively correlated, what is the inconsistency in $\tilde{\beta}_1$, the slope coefficient in the simple regression of *pctstck* on *funds*?

3 The data set SMOKE contains information on smoking behavior and other variables for a random sample of single adults from the United States. The variable *cigs* is the (average) number of cigarettes smoked per day. Do you think *cigs* has a normal distribution in the U.S. adult population? Explain.

4 In the simple regression model (5.16), under the first four Gauss-Markov assumptions, we showed that estimators of the form (5.17) are consistent for the slope, β_1. Given such an estimator, define an estimator of β_0 by $\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1\bar{x}$. Show that plim $\tilde{\beta}_0 = \beta_0$.

5 The following histogram was created using the variable *score* in the data file ECONMATH. Thirty bins were used to create the histogram, and the height of each cell is the proportion of observations falling within the corresponding interval. The best-fitting normal distribution—that is, using the sample mean and sample standard deviation—has been superimposed on the histogram.



(i) If you use the normal distribution to estimate the probability that *score* exceeds 100, would the answer be zero? Why does your answer contradict the assumption of a normal distribution for *score*?

(ii) Explain what is happening in the left tail of the histogram. Does the normal distribution fit well in the left tail?

6 Consider the equation

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$
$$\mathrm{E}(u|x) = 0$$

where the explanatory variable x has a standard normal distribution in the population. In particular, $\mathrm{E}(x) = 0$, $\mathrm{E}(x^2) = Var(x) = 1$, and $\mathrm{E}(x^3) = 0$. This last condition holds because the standard normal distribution is symmetric about zero. We want to study what we can say about the OLS estimator of β_1 we omit x^2 and compute the simple regression estimator of the intercept and slope.

(i) Show that we can write

$$y = \alpha_0 + \beta_1 x + v$$

where $\mathrm{E}(v) = 0$. In particular, find v and the new intercept, α_0.

(ii) Show that $\mathrm{E}(v|x)$ depends on x unless $\beta_2 = 0$.

(iii) Show that $\mathrm{Cov}(x, v) = 0$.

(iv) If $\hat{\beta}_1$ is the slope coefficient from regression y_i on x_i, is $\hat{\beta}_1$ consistent for β_1? Is it unbiased? Explain.

(v) Argue that being able to estimate β_1 has some value in the following sense: β_1 is the partial effect of x on y evaluated at $x = 0$, the average value of x.

(vi) Explain why being able to consistently estimate β_1 and β_2 is more valuable than just estimating β_1.

Computer Exercises

C1 Use the data in WAGE1 for this exercise.

(i) Estimate the equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

Save the residuals and plot a histogram.

(ii) Repeat part (i), but with log(*wage*) as the dependent variable.

(iii) Would you say that Assumption MLR.6 is closer to being satisfied for the level-level model or the log-level model?

C2 Use the data in GPA2 for this exercise.

(i) Using all 4,137 observations, estimate the equation

$$colgpa = \beta_0 + \beta_1 hsperc + \beta_2 sat + u$$

and report the results in standard form.

(ii) Reestimate the equation in part (i), using the first 2,070 observations.

(iii) Find the ratio of the standard errors on *hsperc* from parts (i) and (ii). Compare this with the result from (5.10).

C3 In equation (4.42) of Chapter 4, using the data set BWGHT, compute the *LM* statistic for testing whether *motheduc* and *fatheduc* are jointly significant. In obtaining the residuals for the restricted model, be sure that the restricted model is estimated using only those observations for which all variables in the unrestricted model are available (see Example 4.9).

C4 Several statistics are commonly used to detect nonnormality in underlying population distributions. Here we will study one that measures the amount of skewness in a distribution. Recall that any normally distributed random variable is symmetric about its mean; therefore, if we standardize a symmetrically distributed random variable, say $z = (y - \mu_y)/\sigma_y$, where $\mu_y = \mathrm{E}(y)$ and $\sigma_y = \mathrm{sd}(y)$, then z has mean zero, variance one, and $\mathrm{E}(z^3) = 0$. Given a sample of data $\{y_i: i = 1, \ldots, n\}$, we can standardize y_i in the sample by using $z_i = (y_i - \hat{\mu}_y)/\hat{\sigma}_y$, where $\hat{\mu}_y$ is the sample mean and $\hat{\sigma}_y$ is the sample standard deviation. (We ignore the fact that these are estimates based on the sample.) A sample statistic that measures skewness is $n^{-1}\sum_{i=1}^{n} z_i^3$, or where n is replaced with $(n - 1)$ as a degrees-of-freedom adjustment. If y has a normal distribution in the population, the skewness measure in the sample for the standardized values should not differ significantly from zero.

(i) First use the data set 401KSUBS, keeping only observations with *fsize* = 1. Find the skewness measure for *inc*. Do the same for log(*inc*). Which variable has more skewness and therefore seems less likely to be normally distributed?

(ii) Next use BWGHT2. Find the skewness measures for *bwght* and log(*bwght*). What do you conclude?

(iii) Evaluate the following statement: "The logarithmic transformation always makes a positive variable look more normally distributed."

(iv) If we are interested in the normality assumption in the context of regression, should we be evaluating the unconditional distributions of y and log(y)? Explain.

C5 Consider the analysis in Computer Exercise C11 in Chapter 4 using the data in HTV, where *educ* is the dependent variable in a regression.

(i) How many different values are taken on by *educ* in the sample? Does *educ* have a continuous distribution?

(ii) Plot a histogram of *educ* with a normal distribution overlay. Does the distribution of *educ* appear anything close to normal?

(iii) Which of the CLM assumptions seems clearly violated in the model

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u?$$

How does this violation change the statistical inference procedures carried out in Computer Exercise C11 in Chapter 4?

C6 Use the data in ECONMATH to answer this question.

(i) Logically, what are the smallest and largest values that can be taken on by the variable *score*? What are the smallest and largest values in the sample?

(ii) Consider the linear model

$$score = \beta_0 + \beta_1 colgpa + \beta_2 actmth + \beta_3 acteng + u.$$

Why cannot Assumption MLR.6 hold for the error term u? What consequences does this have for using the usual t statistic to test H_0: $\beta_3 = 0$?

(iii) Estimate the model from part (ii) and obtain the t statistic and associated p-value for testing H_0: $\beta_3 = 0$. How would you defend your findings to someone who makes the following statement: "You cannot trust that p-value because clearly the error term in the equation cannot have a normal distribution."

APPENDIX 5A

Asymptotic Normality of OLS

We sketch a proof of the asymptotic normality of OLS [Theorem 5.2(i)] in the simple regression case. Write the simple regression model as in equation (5.16). Then, by the usual algebra of simple regression, we can write

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1/s_x^2)\left[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right],$$

where we use s_x^2 to denote the sample variance of $\{x_i: i = 1, 2, \ldots, n\}$. By the law of large numbers (see Math Refresher C), $s_x^2 \xrightarrow{p} \sigma_x^2 = \text{Var}(x)$. Assumption MLR.3 rules out perfect collinearity, which means that $\text{Var}(x) > 0$ (x_i varies in the sample, and therefore x is not constant in the population). Next, $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i = n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i + (\mu - \bar{x})[n^{-1/2}\sum_{i=1}^{n}u_i]$, where $\mu = \text{E}(x)$ is the population mean of x. Now $\{u_i\}$ is a sequence of i.i.d. random variables with mean zero and variance σ^2, and so $n^{-1/2}\sum_{i=1}^{n}u_i$ converges to the Normal(0,σ^2) distribution as $n \to \infty$; this is just the central limit theorem from Math Refresher C. By the law of large numbers, $\text{plim}(u - \bar{x}) = 0$. A standard result in asymptotic theory is that if $\text{plim}(w_n) = 0$ and z_n has an asymptotic normal distribution, then $\text{plim}(w_n z_n) = 0$. [See Wooldridge (2010, Chapter 3) for more discussion.] This implies that $(\mu - \bar{x})[n^{-1/2}\sum_{i=1}^{n}u_i]$ has zero plim. Next, $\{(x_i - \mu)u_i: i = 1, 2, \ldots\}$ is an indefinite sequence of i.i.d. random variables with mean zero—because u and x are uncorrelated under Assumption MLR.4—and variance $\sigma^2\sigma_x^2$ by the homoskedasticity Assumption MLR.5. Therefore,

$n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ has an asymptotic Normal$(0,\sigma^2\sigma_x^2)$ distribution. We just showed that the difference between $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ and $n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ has zero plim. A result in asymptotic theory is that if z_n has an asymptotic normal distribution and plim$(v_n - z_n) = 0$, then v_n has the same asymptotic normal distribution as z_n. It follows that $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ also has an asymptotic Normal$(0,\sigma^2\sigma_x^2)$ distribution. Putting all of the pieces together gives

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1/\sigma_x^2)\left[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right] + [(1/s_x^2) - (1/\sigma_x^2)]\left[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right],$$

and because plim$(1/s_x^2) = 1/\sigma_x^2$, the second term has zero plim. Therefore, the asymptotic distribution of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ is Normal$(0,\{\sigma^2\sigma_x^2\}/\{\sigma_x^2\}^2) =$ Normal$(0,\sigma^2/\sigma_x^2)$. This completes the proof in the simple regression case, as $a_1^2 = \sigma_x^2$ in this case. See Wooldridge (2010, Chapter 4) for the general case.

CHAPTER 6

Multiple Regression Analysis: Further Issues

This chapter brings together several issues in multiple regression analysis that we could not conveniently cover in earlier chapters. These topics are not as fundamental as the material in Chapters 3 and 4, but they are important for applying multiple regression to a broad range of empirical problems.

6-1 Effects of Data Scaling on OLS Statistics

In Chapter 2 on bivariate regression, we briefly discussed the effects of changing the units of measurement on the OLS intercept and slope estimates. We also showed that changing the units of measurement did not affect R-squared. We now return to the issue of data scaling and examine the effects of rescaling the dependent or independent variables on standard errors, t statistics, F statistics, and confidence intervals.

We will discover that everything we expect to happen, does happen. When variables are rescaled, the coefficients, standard errors, confidence intervals, t statistics, and F statistics change in ways that preserve all measured effects and testing outcomes. Although this is no great surprise—in fact, we would be very worried if it were not the case—it is useful to see what occurs explicitly. Often, data scaling is used for cosmetic purposes, such as to reduce the number of zeros after a decimal point in an estimated coefficient. By judiciously choosing units of measurement, we can improve the appearance of an estimated equation while changing nothing that is essential.

We could treat this problem in a general way, but it is much better illustrated with examples. Likewise, there is little value here in introducing an abstract notation.

We begin with an equation relating infant birth weight to cigarette smoking and family income:

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc, \qquad [6.1]$$

where

bwght = child birth weight, in ounces.
cigs = number of cigarettes smoked by the mother while pregnant, per day.
faminc = annual family income, in thousands of dollars.

The estimates of this equation, obtained using the data in BWGHT, are given in the first column of Table 6.1. Standard errors are listed in parentheses. The estimate on *cigs* says that if a woman smoked five more cigarettes per day, birth weight is predicted to be about $.4634(5) = 2.317$ ounces less. The *t* statistic on *cigs* is −5.06, so the variable is very statistically significant.

Now, suppose that we decide to measure birth weight in pounds, rather than in ounces. Let $bwghtlbs = bwght/16$ be birth weight in pounds. What happens to our OLS statistics if we use this as the dependent variable in our equation? It is easy to find the effect on the coefficient estimates by simple manipulation of equation (6.1). Divide this entire equation by 16:

$$\widehat{bwght}/16 = \hat{\beta}_0/16 + (\hat{\beta}_1/16)cigs + (\hat{\beta}_2/16)faminc.$$

Because the left-hand side is birth weight in pounds, it follows that each new coefficient will be the corresponding old coefficient divided by 16. To verify this, the regression of *bwghtlbs* on *cigs*, and *faminc* is reported in column (2) of Table 6.1. Up to the reported digits (and any digits beyond), the intercept and slopes in column (2) are just those in column (1) divided by 16. For example, the coefficient on *cigs* is now −.0289; this means that if *cigs* were higher by five, birth weight would be $.0289(5) = .1445$ *pounds* lower. In terms of ounces, we have $.1445(16) = 2.312$, which is slightly different from the 2.317 we obtained earlier due to rounding error. The point is, after the effects are transformed into the same units, we get exactly the same answer, regardless of how the dependent variable is measured.

What about statistical significance? As we expect, changing the dependent variable from ounces to pounds has no effect on how statistically important the independent variables are. The standard errors in column (2) are 16 times smaller than those in column (1). A few quick calculations show

TABLE 6.1 Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Independent Variables			
cigs	−.4634 (.0916)	−.0289 (.0057)	—
packs	—	—	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercept	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observations	1,388	1,388	1,388
R-Squared	.0298	.0298	.0298
SSR	557,485.51	2,177.6778	557,485.51
SER	20.063	1.2539	20.063

that the t statistics in column (2) are indeed identical to the t statistics in column (1). The endpoints for the confidence intervals in column (2) are just the endpoints in column (1) divided by 16. This is because the CIs change by the same factor as the standard errors. [Remember that the 95% CI here is $\hat{\beta}_j \pm 1.96\,\text{se}(\hat{\beta}_j)$.]

In terms of goodness-of-fit, the R-squareds from the two regressions are identical, as should be the case. Notice that the sum of squared residuals (SSR) and the standard error of the regression (SER) do differ across equations. These differences are easily explained. Let $\hat{u}_i$ denote the residual for observation i in the original equation (6.1). Then the residual when *bwghtlbs* is the dependent variable is simply $\hat{u}_i/16$. Thus, the *squared* residual in the second equation is $(\hat{u}_i/16)^2 = \hat{u}_i^2/256$. This is why the SSR in column (2) is equal to the SSR in column (1) divided by 256.

Because $\text{SER} = \hat{\sigma} = \sqrt{\text{SSR}/(n-k-1)} = \sqrt{\text{SSR}/1{,}385}$, the SER in column (2) is 16 times smaller than that in column (1). Another way to think about this is that the error in the equation with *bwghtlbs* as the dependent variable has a standard deviation 16 times smaller than the standard deviation of the original error. This does not mean that we have reduced the error by changing how birth weight is measured; the smaller SER simply reflects a difference in units of measurement.

Next, let us return the dependent variable to its original units: *bwght* is measured in ounces. Instead, let us change the unit of measurement of one of the independent variables, *cigs*. Define *packs* to be the number of packs of cigarettes smoked per day. Thus, $packs = cigs/20$. What happens to the coefficients and other OLS statistics now? Well, we can write

$$\widehat{bwght} = \hat{\beta}_0 + (20\hat{\beta}_1)(cigs/20) + \hat{\beta}_2\, faminc = \hat{\beta}_0 + (20\hat{\beta}_1)packs + \hat{\beta}_2\, faminc.$$

Thus, the intercept and slope coefficient on *faminc* are unchanged, but the coefficient on *packs* is 20 times that on *cigs*. This is intuitively appealing. The results from the regression of *bwght* on *packs* and *faminc* are in column (3) of Table 6.1. Incidentally, remember that it would make no sense to include both *cigs* and *packs* in the same equation; this would induce perfect multicollinearity and would have no interesting meaning.

GOING FURTHER 6.1

In the original birth weight equation (6.1), suppose that *faminc* is measured in dollars rather than in thousands of dollars. Thus, define the variable *fincdol* = 1,000·*faminc*. How will the OLS statistics change when *fincdol* is substituted for *faminc*? For the purpose of presenting the regression results, do you think it is better to measure income in dollars or in thousands of dollars?

Other than the coefficient on *packs*, there is one other statistic in column (3) that differs from that in column (1): the standard error on *packs* is 20 times larger than that on *cigs* in column (1). This means that the t statistic for testing the significance of cigarette smoking is the same whether we measure smoking in terms of cigarettes or packs. This is only natural.

The previous example spells out most of the possibilities that arise when the dependent and independent variables are rescaled. Rescaling is often done with dollar amounts in economics, especially when the dollar amounts are very large.

In Chapter 2, we argued that, if the dependent variable appears in logarithmic form, changing the unit of measurement does not affect the slope coefficient. The same is true here: changing the unit of measurement of the dependent variable, when it appears in logarithmic form, does not affect any of the slope estimates. This follows from the simple fact that $\log(c_1 y_i) = \log(c_1) + \log(y_i)$ for any constant $c_1 > 0$. The new intercept will be $\log(c_1) + \hat{\beta}_0$. Similarly, changing the unit of measurement of any x_j, where $\log(x_j)$ appears in the regression, only affects the intercept. This corresponds to what we know about percentage changes and, in particular, elasticities: they are invariant to the units of measurement of either y or the x_j. For example, if we had specified the dependent variable in (6.1) to be $\log(bwght)$, estimated the equation, and then reestimated it with $\log(bwghtlbs)$ as the dependent variable, the coefficients on *cigs* and *faminc* would be the same in both regressions; only the intercept would be different.

6-1a Beta Coefficients

Sometimes, in econometric applications, a key variable is measured on a scale that is difficult to interpret. Labor economists often include test scores in wage equations, and the scale on which these tests are scored is often arbitrary and not easy to interpret (at least for economists!). In almost all cases, we are interested in how a particular individual's score compares with the population. Thus, instead of asking about the effect on hourly wage if, say, a test score is 10 points higher, it makes more sense to ask what happens when the test score is one *standard deviation* higher.

Nothing prevents us from seeing what happens to the dependent variable when an independent variable in an estimated model increases by a certain number of standard deviations, assuming that we have obtained the sample standard deviation of the independent variable (which is easy in most regression packages). This is often a good idea. So, for example, when we look at the effect of a standardized test score, such as the SAT score, on college GPA, we can find the standard deviation of SAT and see what happens when the SAT score increases by one or two standard deviations.

Sometimes, it is useful to obtain regression results when *all* variables involved, the dependent as well as all the independent variables, have been *standardized*. A variable is standardized in the sample by subtracting off its mean and dividing by its standard deviation (see Math Refresher C). This means that we compute the *z-score* for every variable in the sample. Then, we run a regression using the z-scores.

Why is standardization useful? It is easiest to start with the original OLS equation, with the variables in their original forms:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + \hat{u}_i. \qquad [6.2]$$

We have included the observation subscript i to emphasize that our standardization is applied to all sample values. Now, if we average (6.2), use the fact that the $\hat{u}_i$ have a zero sample average, and subtract the result from (6.2), we get

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \cdots + \hat{\beta}_k(x_{ik} - \bar{x}_k) + \hat{u}_i.$$

Now, let $\hat{\sigma}_y$ be the sample standard deviation for the dependent variable, let $\hat{\sigma}_1$ be the sample *sd* for x_1, let $\hat{\sigma}_2$ be the sample *sd* for x_2, and so on. Then, simple algebra gives the equation

$$\begin{aligned}(y_i - \bar{y})/\hat{\sigma}_y = (\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1[(x_{i1} - \bar{x}_1)/\hat{\sigma}_1] + \cdots \\ + (\hat{\sigma}_k/\hat{\sigma}_y)\hat{\beta}_k[(x_{ik} - \bar{x}_k)/\hat{\sigma}_k] + (\hat{u}_i/\hat{\sigma}_y).\end{aligned} \qquad [6.3]$$

Each variable in (6.3) has been standardized by replacing it with its z-score, and this has resulted in new slope coefficients. For example, the slope coefficient on $(x_{i1} - \bar{x}_1)/\hat{\sigma}_1$ is $(\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1$. This is simply the original coefficient, $\hat{\beta}_1$, multiplied by the ratio of the standard deviation of x_1 to the standard deviation of y. The intercept has dropped out altogether.

It is useful to rewrite (6.3), dropping the i subscript, as

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \cdots + \hat{b}_k z_k + error, \qquad [6.4]$$

where z_y denotes the z-score of y, z_1 is the z-score of x_1, and so on. The new coefficients are

$$\hat{b}_j = (\hat{\sigma}_j/\hat{\sigma}_y)\hat{\beta}_j \text{ for } j = 1, \ldots, k. \qquad [6.5]$$

These $\hat{b}_j$ are traditionally called **standardized coefficients** or **beta coefficients**. (The latter name is more common, which is unfortunate because we have been using beta hat to denote the *usual* OLS estimates.)

Beta coefficients receive their interesting meaning from equation (6.4): if x_1 increases by one standard deviation, then $\hat{y}$ changes by $\hat{b}_1$ standard deviations. Thus, we are measuring effects not in terms of the original units of y or the x_j, but in standard deviation units. Because it makes the scale of the regressors irrelevant, this equation puts the explanatory variables on equal footing. In a standard OLS equation, it is not possible to simply look at the size of different coefficients and conclude that the explanatory variable with the largest coefficient is "the most important." We just saw that the magnitudes of coefficients can be changed at will by changing the units of measurement of the x_j. But, when each x_j has been standardized, comparing the magnitudes of the resulting beta coefficients is more compelling. When the regression equation has only a single explanatory variable, x_1, its standardized coefficient is simply the sample correlation coefficient between y and x_1, which means it must lie in the range -1 to 1.

Even in situations in which the coefficients are easily interpretable—say, the dependent variable and independent variables of interest are in logarithmic form, so the OLS coefficients of interest are estimated elasticities—there is still room for computing beta coefficients. Although elasticities are free of units of measurement, a change in a particular explanatory variable by, say, 10% may represent a larger or smaller change over a variable's range than changing another explanatory variable by 10%. For example, in a state with wide income variation but relatively little variation in spending per student, it might not make much sense to compare performance elasticities with respect to the income and spending. Comparing beta coefficient magnitudes can be helpful.

To obtain the beta coefficients, we can always standardize $y, x_1, \ldots, x_k$ and then run the OLS regression of the z-score of y on the z-scores of $x_1, \ldots, x_k$—where it is not necessary to include an intercept, as it will be zero. This can be tedious with many independent variables. Many regression packages provide beta coefficients via a simple command. The following example illustrates the use of beta coefficients.

EXAMPLE 6.1 Effects of Pollution on Housing Prices

We use the data from Example 4.5 (in the file HPRICE2) to illustrate the use of beta coefficients. Recall that the key independent variable is *nox*, a measure of the nitrogen oxide in the air over each community. One way to understand the size of the pollution effect—without getting into the science underlying nitrogen oxide's effect on air quality—is to compute beta coefficients. (An alternative approach is contained in Example 4.5: we obtained a price elasticity with respect to *nox* by using *price* and *nox* in logarithmic form.)

The population equation is the level-level model

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + u,$$

where all the variables except *crime* were defined in Example 4.5; *crime* is the number of reported crimes per capita. The beta coefficients are reported in the following equation (so each variable has been converted to its z-score):

$$\widehat{zprice} = -.340\, znox - .143\, zcrime + .514\, zrooms - .235\, zdist - .270\, zstratio.$$

This equation shows that a one standard deviation increase in *nox* decreases price by .34 standard deviation; a one standard deviation increase in *crime* reduces price by .14 standard deviation. Thus, the same relative movement of pollution in the population has a larger effect on housing prices than crime does. Size of the house, as measured by number of rooms (*rooms*), has the largest standardized effect. If we want to know the effects of each independent variable on the dollar value of median house price, we should use the unstandardized variables.

Whether we use standardized or unstandardized variables does not affect statistical significance: the t statistics are the same in both cases.

6-2 More on Functional Form

In several previous examples, we have encountered the most popular device in econometrics for allowing nonlinear relationships between the explained and explanatory variables: using logarithms for the dependent or independent variables. We have also seen models containing quadratics in some explanatory variables, but we have yet to provide a systematic treatment of them. In this section, we cover some variations and extensions on functional forms that often arise in applied work.

6-2a More on Using Logarithmic Functional Forms

We begin by reviewing how to interpret the parameters in the model

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u, \qquad [6.6]$$

where these variables are taken from Example 4.5. Recall that throughout the text $\log(x)$ is the *natural* log of x. The coefficient β_1 is the elasticity of *price* with respect to *nox* (pollution). The coefficient β_2 is the change in $\log(price)$, when $\Delta rooms = 1$; as we have seen many times, when multiplied by 100, this is the approximate percentage change in *price*. Recall that $100 \cdot \beta_2$ is sometimes called the semi-elasticity of *price* with respect to *rooms*.

When estimated using the data in HPRICE2, we obtain

$$\widehat{\log(price)} = 9.23 - .718 \log(nox) + .306\, rooms$$
$$(0.19) \quad (.066) \qquad\qquad (.019) \qquad [6.7]$$
$$n = 506, R^2 = .514.$$

Thus, when *nox* increases by 1%, *price* falls by .718%, holding only *rooms* fixed. When *rooms* increases by one, *price* increases by approximately $100(.306) = 30.6\%$.

The estimate that one more room increases price by about 30.6% turns out to be somewhat inaccurate for this application. The approximation error occurs because, as the change in $\log(y)$ becomes larger and larger, the approximation $\%\Delta y \approx 100 \cdot \Delta \log(y)$ becomes more and more inaccurate. Fortunately, a simple calculation is available to compute the exact percentage change.

To describe the procedure, we consider the general estimated model

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 x_2.$$

(Adding additional independent variables does not change the procedure.) Now, fixing x_1, we have $\Delta\widehat{\log(y)} = \hat{\beta}_2 \Delta x_2$. Using simple algebraic properties of the exponential and logarithmic functions gives the *exact* percentage change in the predicted y as

$$\%\Delta\hat{y} = 100 \cdot [\exp(\hat{\beta}_2 \Delta x_2) - 1], \qquad [6.8]$$

where the multiplication by 100 turns the proportionate change into a percentage change. When $\Delta x_2 = 1$,

$$\%\Delta\hat{y} = 100 \cdot [\exp(\hat{\beta}_2) - 1]. \qquad [6.9]$$

Applied to the housing price example with $x_2 = rooms$ and $\hat{\beta}_2 = .306$, $\%\Delta\widehat{price} = 100[\exp(.306) - 1] = 35.8\%$, which is notably larger than the approximate percentage change, 30.6%, obtained directly from (6.7). {Incidentally, this is not an unbiased estimator because $\exp(\cdot)$ is a nonlinear function; it is, however, a consistent estimator of $100[\exp(\beta_2) - 1]$. This is because the probability limit passes through continuous functions, while the expected value operator does not. See Math Refresher C.}

The adjustment in equation (6.8) is not as crucial for small percentage changes. For example, when we include the student-teacher ratio in equation (6.7), its estimated coefficient is –.052, which means that if *stratio* increases by one, *price* decreases by approximately 5.2%. The exact proportionate change is $\exp(-.052)-1 \approx -.051$, or -5.1%. On the other hand, if we increase *stratio* by five, then the approximate percentage change in price is –26%, while the exact change obtained from equation (6.8) is $100[\exp(-.26)-1] \approx -22.9\%$.

The logarithmic approximation to percentage changes has an advantage that justifies its reporting even when the percentage change is large. To describe this advantage, consider again the effect on price of changing the number of rooms by one. The logarithmic approximation is just the coefficient on rooms in equation (6.7) multiplied by 100, namely, 30.6%. We also computed an estimate of the exact percentage change for *increasing* the number of rooms by one as 35.8%. But what if we want to estimate the percentage change for *decreasing* the number of rooms by one? In equation (6.8) we take $\Delta x_2 = -1$ and $\hat{\beta}_2 = .306$, and so $\%\Delta\widehat{price} = 100[\exp(-.306)-1] = -26.4$, or a drop of 26.4%. Notice that the approximation based on using the coefficient on *rooms* is between 26.4 and 35.8—an outcome that always occurs. In other words, simply using the coefficient (multiplied by 100) gives us an estimate that is always between the absolute value of the estimates for an increase and a decrease. If we are specifically interested in an increase or a decrease, we can use the calculation based on equation (6.8).

The point just made about computing percentage changes is essentially the one made in introductory economics when it comes to computing, say, price elasticities of demand based on large price changes: the result depends on whether we use the beginning or ending price and quantity in computing the percentage changes. Using the logarithmic approximation is similar in spirit to calculating an arc elasticity of demand, where the averages of prices and quantities are used in the denominators in computing the percentage changes.

We have seen that using natural logs leads to coefficients with appealing interpretations, and we can be ignorant about the units of measurement of variables appearing in logarithmic form because the slope coefficients are invariant to rescalings. There are several other reasons logs are used so much in applied work. First, when $y > 0$, models using $\log(y)$ as the dependent variable often satisfy the CLM assumptions more closely than models using the level of y. Strictly positive variables often have conditional distributions that are heteroskedastic or skewed; taking the log can mitigate, if not eliminate, both problems.

Another potential benefit of using logs is that taking the log of a variable often narrows its range. This is particularly true of variables that can be large monetary values, such as firms' annual sales or baseball players' salaries. Population variables also tend to vary widely. Narrowing the range of the dependent and independent variables can make OLS estimates less sensitive to outlying (or extreme) values; we take up the issue of outlying observations in Chapter 9.

However, one must not indiscriminately use the logarithmic transformation because in some cases it can actually create extreme values. An example is when a variable y is between zero and one (such as a proportion) and takes on values close to zero. In this case, $\log(y)$ (which is necessarily negative) can be very large in magnitude whereas the original variable, y, is bounded between zero and one.

There are some standard rules of thumb for taking logs, although none is written in stone. When a variable is a positive dollar amount, the log is often taken. We have seen this for variables such as wages, salaries, firm sales, and firm market value. Variables such as population, total number of employees, and school enrollment often appear in logarithmic form; these have the common feature of being large integer values.

Variables that are measured in years—such as education, experience, tenure, age, and so on—usually appear in their original form. A variable that is a proportion or a percent—such as the unemployment rate, the participation rate in a pension plan, the percentage of students passing a standardized exam, and the arrest rate on reported crimes—can appear in either original or logarithmic form, although there is a tendency to use them in level forms. This is because any regression

coefficients involving the *original* variable—whether it is the dependent or independent variable—will have a *percentage point* change interpretation. (See Math Refresher A for a review of the distinction between a percentage change and a percentage point change.) If we use, say, log(*unem*) in a regression, where *unem* is the percentage of unemployed individuals, we must be very careful to distinguish between a percentage point change and a percentage change. Remember, if *unem* goes from 8 to 9, this is an increase of one percentage point, but a 12.5% increase from the initial unemployment level. Using the log means that we are looking at the percentage change in the unemployment rate: $\log(9) - \log(8) \approx .118$ or 11.8%, which is the logarithmic approximation to the actual 12.5% increase.

One limitation of the log is that it cannot be used if a variable takes on zero or negative values. In cases where a variable y is nonnegative but can take on the value 0, $\log(1+y)$ is sometimes used. The percentage change interpretations are often closely preserved, except for changes beginning at $y = 0$ (where the percentage change is not even defined). Generally, using $\log(1+y)$ and then interpreting the estimates as if the variable were $\log(y)$ is acceptable when the data on y contain relatively few zeros. An example might be where y is hours of training per employee for the population of manufacturing firms, if a large fraction of firms provides training to at least one worker. Technically, however, $\log(1+y)$ cannot be normally distributed (although it might be less heteroskedastic than y). Useful, albeit more advanced, alternatives are the Tobit and Poisson models in Chapter 17.

One drawback to using a dependent variable in logarithmic form is that it is more difficult to predict the original variable. The original model allows us to predict $\log(y)$, not y. Nevertheless, it is fairly easy to turn a prediction for $\log(y)$ into a prediction for y (see Section 6-4). A related point is that it is *not* legitimate to compare R-squareds from models where y is the dependent variable in one case and $\log(y)$ is the dependent variable in the other. These measures explain variations in different variables. We discuss how to compute comparable goodness-of-fit measures in Section 6-4.

GOING FURTHER 6.2

Suppose that the annual number of drunk driving arrests is determined by

$$\log(arrests) = \beta_0 + \beta_1 \log(pop) + \beta_2 age16\_25 + other\ factors,$$

where *age 16_25* is the proportion of the population between 16 and 25 years of age. Show that β_2 has the following (ceteris paribus) interpretation: it is the percentage change in *arrests* when the percentage of the people aged 16 to 25 increases by one *percentage point*.

6-2b Models with Quadratics

Quadratic functions are also used quite often in applied economics to capture decreasing or increasing marginal effects. You may want to review properties of quadratic functions in Math Refresher A.

In the simplest case, y depends on a single observed factor x, but it does so in a quadratic fashion:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

For example, take $y = wage$ and $x = exper$. As we discussed in Chapter 3, this model falls outside of simple regression analysis but is easily handled with multiple regression.

It is important to remember that β_1 does not measure the change in y with respect to x; it makes no sense to hold x^2 fixed while changing x. If we write the estimated equation as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2, \qquad [6.10]$$

then we have the approximation

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x, \text{ so } \Delta\hat{y}/\Delta x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x. \qquad [6.11]$$

This says that the slope of the relationship between x and y depends on the value of x; the estimated slope is $\hat{\beta}_1 + 2\hat{\beta}_2 x$. If we plug in $x = 0$, we see that $\hat{\beta}_1$ can be interpreted as the approximate slope in going from $x = 0$ to $x = 1$. After that, the second term, $2\hat{\beta}_2 x$, must be accounted for.

If we are only interested in computing the predicted change in y given a starting value for x and a change in x, we could use (6.10) directly: there is no reason to use the calculus approximation at all. However, we are usually more interested in quickly summarizing the effect of x on y, and the interpretation of $\hat{\beta}_1$ and $\hat{\beta}_2$ in equation (6.11) provides that summary. Typically, we might plug in the average value of x in the sample, or some other interesting values, such as the median or the lower and upper quartile values.

In many applications, $\hat{\beta}_1$ is positive and $\hat{\beta}_2$ is negative. For example, using the wage data in WAGE1, we obtain

$$\widehat{wage} = 3.73 + .298\, exper - .0061\, exper^2 \quad [6.12]$$
$$(.35)\ \ (.041)\ \ \ \ \ \ (.0009)$$
$$n = 526,\ R^2 = .093.$$

This estimated equation implies that *exper* has a diminishing effect on *wage*. The first year of experience is worth roughly 30¢ per hour ($.298). The second year of experience is worth less [about $.298 - 2(.0061)(1) \approx .286$, or 28.6¢, according to the approximation in (6.11) with $x = 1$]. In going from 10 to 11 years of experience, *wage* is predicted to increase by about $.298 - 2(.0061)(10) = .176$, or 17.6¢. And so on.

When the coefficient on x is positive and the coefficient on x^2 is negative, the quadratic has a parabolic shape. There is always a positive value of x where the effect of x on y is zero; before this point, x has a positive effect on y; after this point, x has a negative effect on y. In practice, it can be important to know where this turning point is.

In the estimated equation (6.10) with $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$, the turning point (or maximum of the function) is always achieved at the coefficient on x over *twice* the absolute value of the coefficient on x^2:

$$x^* = |\hat{\beta}_1/(2\hat{\beta}_2)|. \quad [6.13]$$

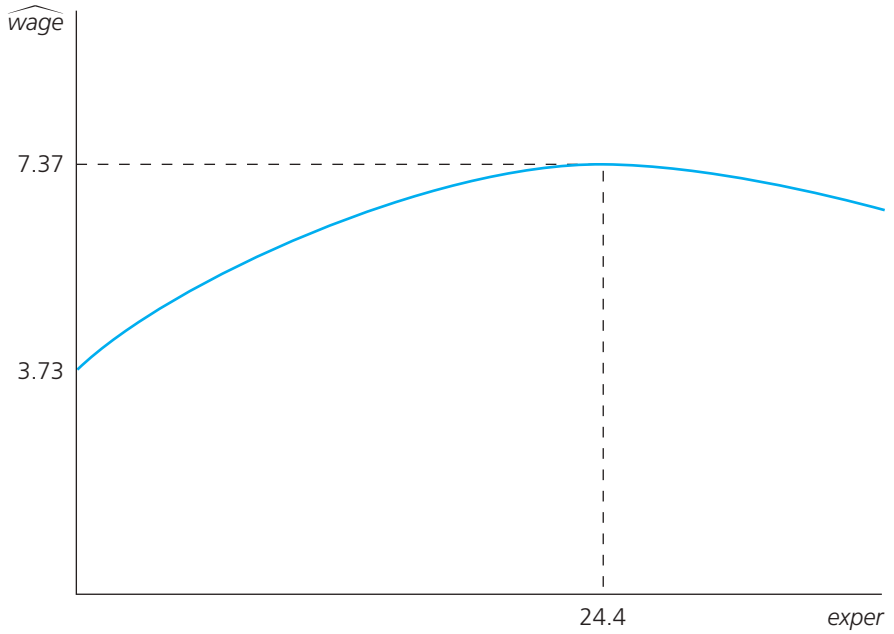
In the wage example, $x^* = exper^*$ is $.298/[2(.0061)] \approx 24.4$. (Note how we just drop the minus sign on −.0061 in doing this calculation.) This quadratic relationship is illustrated in Figure 6.1.

In the wage equation (6.12), the return to experience becomes zero at about 24.4 years. What should we make of this? There are at least three possible explanations. First, it may be that few people in the sample have more than 24 years of experience, and so the part of the curve to the right of 24 can be ignored. The cost of using a quadratic to capture diminishing effects is that the quadratic must eventually turn around. If this point is beyond all but a small percentage of the people in the sample, then this is not of much concern. But in the data set WAGE1, about 28% of the people in the sample have more than 24 years of experience; this is too high a percentage to ignore.

It is possible that the return to *exper* really becomes negative at some point, but it is hard to believe that this happens at 24 years of experience. A more likely possibility is that the estimated effect of *exper* on *wage* is biased because we have controlled for no other factors, or because the functional relationship between *wage* and *exper* in equation (6.12) is not entirely correct. Computer Exercise C2 asks you to explore this possibility by controlling for education, in addition to using log(*wage*) as the dependent variable.

When a model has a dependent variable in logarithmic form and an explanatory variable entering as a quadratic, some care is needed in reporting the partial effects. The following example also shows that the quadratic can have a U-shape, rather than a parabolic shape. A U-shape arises in equation (6.10) when $\hat{\beta}_1$ is negative and $\hat{\beta}_2$ is positive; this captures an increasing effect of x on y.

FIGURE 6.1 Quadratic relationship between $\widehat{wage}$ and *exper*.



EXAMPLE 6.2 Effects of Pollution on Housing Prices

We modify the housing price model from Example 4.5 to include a quadratic term in *rooms*:

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + u. \quad [6.14]$$

The model estimated using the data in HPRICE2 is

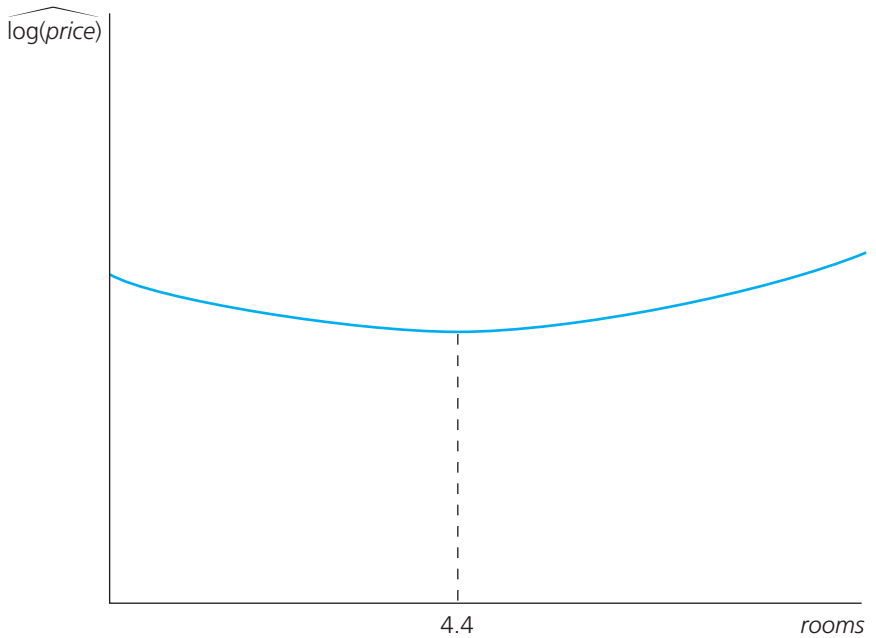
$$\widehat{\log(price)} = \underset{(.57)}{13.39} - \underset{(.115)}{.902}\log(nox) - \underset{(.043)}{.087}\log(dist) - \underset{(.165)}{.545}\, rooms + \underset{(.013)}{.062}\, rooms^2 - \underset{(.006)}{.048}\, stratio$$
$$n = 506,\ R^2 = .603.$$

The quadratic term $rooms^2$ has a t statistic of about 4.77, and so it is very statistically significant. But what about interpreting the effect of *rooms* on log(*price*)? Initially, the effect appears to be strange. Because the coefficient on *rooms* is negative and the coefficient on $rooms^2$ is positive, this equation literally implies that, at low values of *rooms*, an additional room has a *negative* effect on log(*price*). At some point, the effect becomes positive, and the quadratic shape means that the semi-elasticity of *price* with respect to *rooms* is increasing as *rooms* increases. This situation is shown in Figure 6.2.

We obtain the turnaround value of *rooms* using equation (6.13) (even though $\hat{\beta}_1$ is negative and $\hat{\beta}_2$ is positive). The absolute value of the coefficient on *rooms*, .545, divided by twice the coefficient on $rooms^2$, .062, gives $rooms^* = .545/[2(.062)] \approx 4.4$; this point is labeled in Figure 6.2.

Do we really believe that starting at three rooms and increasing to four rooms actually reduces a house's expected value? Probably not. It turns out that only five of the 506 communities in the sample

FIGURE 6.2 $\widehat{\log(price)}$ **as a quadratic function of** *rooms*.



have houses averaging 4.4 rooms or less, about 1% of the sample. This is so small that the quadratic to the left of 4.4 can, for practical purposes, be ignored. To the right of 4.4, we see that adding another room has an increasing effect on the percentage change in price:

$$\Delta\widehat{\log(price)} \approx \{[-.545 + 2(.062)]rooms\}\Delta rooms$$

and so

$$\begin{aligned}\%\Delta\widehat{price} &\approx 100\{[-.545 + 2(.062)]rooms\}\Delta rooms \\ &= (-54.5 + 12.4\, rooms)\Delta rooms.\end{aligned}$$

Thus, an increase in *rooms* from, say, five to six increases price by about $-54.5 + 12.4(5) = 7.5\%$; the increase from six to seven increases price by roughly $-54.5 + 12.4(6) = 19.9\%$. This is a very strong increasing effect.

The strong increasing effect of rooms on log(*price*) in this example illustrates an important lesson: one cannot simply look at the coefficient on the quadratic term—in this case, .062—and declare that it is too small to bother with, based only on its magnitude. In many applications with quadratics, the coefficient on the squared variable has one or more zeros after the decimal point: after all, this coefficient measures how the slope is changing as x (*rooms*) changes. A seemingly small coefficient can have practically important consequences, as we just saw. As a general rule, one must compute the partial effect and see how it varies with x to determine if the quadratic term is practically important. In doing so, it is useful to compare the changing slope implied by the quadratic model with the constant slope obtained from the model with only a linear term. If we drop $rooms^2$ from the equation, the coefficient on *rooms* becomes about .255, which implies that each additional room—starting from any number of rooms—increases median price by about 25.5%. This is very different from the quadratic model, where the effect becomes 25.5% at $rooms = 6.45$ but changes rapidly as *rooms* gets smaller or larger. For example, at $rooms = 7$, the return to the next room is about 32.3%.

What happens generally if the coefficients on the level and squared terms have the *same* sign (either both positive or both negative) and the explanatory variable is necessarily nonnegative (as in the case of *rooms* or *exper*)? In either case, there is no turning point for values $x > 0$. For example, if β_1 and β_2 are both positive, the smallest expected value of y is at $x = 0$, and increases in x always have a positive and increasing effect on y. (This is also true if $\beta_1 = 0$ and $\beta_2 > 0$, which means that the partial effect is zero at $x = 0$ and increasing as x increases.) Similarly, if β_1 and β_2 are both negative, the largest expected value of y is at $x = 0$, and increases in x have a negative effect on y, with the magnitude of the effect increasing as x gets larger.

The general formula for the turning point of any quadratic is $x^* = -\hat{\beta}_1/(2\hat{\beta}_2)$, which leads to a positive value if $\hat{\beta}_1$ and $\hat{\beta}_2$ have opposite signs and a negative value when $\hat{\beta}_1$ and $\hat{\beta}_2$ have the same sign. Knowing this simple formula is useful in cases where x may take on both positive and negative values; one can compute the turning point and see if it makes sense, taking into account the range of x in the sample.

There are many other possibilities for using quadratics along with logarithms. For example, an extension of (6.14) that allows a nonconstant elasticity between *price* and *nox* is

$$\log(price) = \beta_0 + \beta_1\log(nox) + \beta_2[\log(nox)]^2 + \beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u. \quad [6.15]$$

If $\beta_2 = 0$, then β_1 is the elasticity of *price* with respect to *nox*. Otherwise, this elasticity depends on the level of *nox*. To see this, we can combine the arguments for the partial effects in the quadratic and logarithmic models to show that

$$\%\Delta price \approx [\beta_1 + 2\beta_2\log(nox)]\%\Delta nox; \quad [6.16]$$

therefore, the elasticity of *price* with respect to *nox* is $\beta_1 + 2\beta_2\log(nox)$, so that it depends on $\log(nox)$.

Finally, other polynomial terms can be included in regression models. Certainly, the quadratic is seen most often, but a cubic and even a quartic term appear now and then. An often reasonable functional form for a total cost function is

$$cost = \beta_0 + \beta_1 quantity + \beta_2 quantity^2 + \beta_3 quantity^3 + u.$$

Estimating such a model causes no complications. Interpreting the parameters is more involved (though straightforward using calculus); we do not study these models further.

6-2c Models with Interaction Terms

Sometimes, it is natural for the partial effect, elasticity, or semi-elasticity of the dependent variable with respect to an explanatory variable to depend on the magnitude of yet *another* explanatory variable. For example, in the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft\cdot bdrms + \beta_4 bthrms + u,$$

the partial effect of *bdrms* on *price* (holding all other variables fixed) is

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft. \quad [6.17]$$

If $\beta_3 > 0$, then (6.17) implies that an additional bedroom yields a higher increase in housing price for larger houses. In other words, there is an **interaction effect** between square footage and number of bedrooms. In summarizing the effect of *bdrms* on *price*, we must evaluate (6.17) at interesting values

of *sqrft*, such as the mean value, or the lower and upper quartiles in the sample. Whether or not β_3 is zero is something we can easily test.

The parameters on the original variables can be tricky to interpret when we include an interaction term. For example, in the previous housing price equation, equation (6.17) shows that β_2 is the effect of *bdrms* on *price* for a home with zero square feet! This effect is clearly not of much interest. Instead, we must be careful to put interesting values of *sqrft*, such as the mean or median values in the sample, into the estimated version of equation (6.17).

Often, it is useful to reparameterize a model so that the coefficients on the original variables have an interesting meaning. Consider a model with two explanatory variables and an interaction:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

As just mentioned, β_2 is the partial effect of x_2 on y when $x_1 = 0$. Often, this is not of interest. Instead, we can reparameterize the model as

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u,$$

where μ_1 is the population mean of x_1 and μ_2 is the population mean of x_2. We can easily see that now the coefficient on x_2, δ_2, is the partial effect of x_2 on y at the mean value of x_1. (By multiplying out the interaction in the second equation and comparing the coefficients, we can easily show that $\delta_2 = \beta_2 + \beta_3 \mu_1$. The parameter δ_1 has a similar interpretation.) Therefore, if we subtract the means of the variables—in practice, these would typically be the sample means—before creating the interaction term, the coefficients on the original variables have a useful interpretation. Plus, we immediately obtain standard errors for the partial effects at the mean values. Nothing prevents us from replacing μ_1 or μ_2 with other values of the explanatory variables that may be of interest. The following example illustrates how we can use interaction terms.

EXAMPLE 6.3 Effects of Attendance on Final Exam Performance

A model to explain the standardized outcome on a final exam (*stndfnl*) in terms of percentage of classes attended, prior college grade point average, and ACT score is

$$\begin{aligned} stndfnl = {} & \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 \\ & + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + u. \end{aligned} \qquad [6.18]$$

(We use the standardized exam score for the reasons discussed in Section 6-1: it is easier to interpret a student's performance relative to the rest of the class.) In addition to quadratics in *priGPA* and *ACT*, this model includes an interaction between *priGPA* and the attendance rate. The idea is that class attendance might have a different effect for students who have performed differently in the past, as measured by *priGPA*. We are interested in the effects of attendance on final exam score: $\Delta stndfnl/\Delta atndrte = \beta_1 + \beta_6 priGPA$.

Using the 680 observations in ATTEND, for students in a course on microeconomic principles, the estimated equation is

$$\begin{aligned} \widehat{stndfnl} = {} & 2.05 - .0067\, atndrte - 1.63\, priGPA - .128\, ACT \\ & \quad (1.36)\ (.0102) \qquad (.48) \qquad (.098) \\ & + .296\, priGPA^2 + .0045\, ACT^2 + .0056\, priGPA \cdot atndrte \\ & \quad (.101) \qquad (.0022) \qquad (.0043) \\ & n = 680,\ R^2 = .229,\ \overline{R}^2 = .222. \end{aligned} \qquad [6.19]$$

We must interpret this equation with extreme care. If we simply look at the coefficient on *atndrte*, we will incorrectly conclude that attendance has a *negative* effect on final exam score. But this coefficient supposedly measures the effect when $priGPA = 0$, which is not interesting (in this sample, the smallest prior GPA is about .86). We must also take care not to look separately at the estimates of β_1 and β_6 and conclude that, because each t statistic is insignificant, we cannot reject H_0: $\beta_1 = 0, \beta_6 = 0$. In fact, the p-value for the F test of this joint hypothesis is .014, so we certainly reject H_0 at the 5% level. This is a good example of where looking at separate t statistics when testing a joint hypothesis can lead one far astray.

How should we estimate the partial effect of *atndrte* on *stndfnl*? We must plug in interesting values of *priGPA* to obtain the partial effect. The mean value of *priGPA* in the sample is 2.59, so at the mean *priGPA*, the effect of *atndrte* on *stndfnl* is $-.0067 + .0056(2.59) \approx .0078$. What does this mean? Because *atndrte* is measured as a percentage, it means that a 10 percentage point increase in *atndrte* increases $\widehat{stndfnl}$ by .078 standard deviations from the mean final exam score.

GOING FURTHER 6.3

If we add the term $\beta_7\, ACT{\cdot}atndrte$ to equation (6.18), what is the partial effect of *atndrte* on *stndfnl*?

How can we tell whether the estimate .0078 is statistically different from zero? We need to rerun the regression, where we replace $priGPA{\cdot}atndrte$ with $(priGPA - 2.59){\cdot}atndrte$. This gives, as the new coefficient on *atndrte*, the estimated effect at $priGPA = 2.59$, along with its standard error; nothing else in the regression changes. (We described this device in Section 4.4.) Running this new regression gives the standard error of $\hat{\beta}_1 + \hat{\beta}_6(2.59) = .0078$ as .0026, which yields $t = .0078/.0026 = 3$. Therefore, at the average *priGPA*, we conclude that attendance has a statistically significant positive effect on final exam score.

Things are even more complicated for finding the effect of *priGPA* on *stndfnl* because of the quadratic term $priGPA^2$. To find the effect at the mean value of *priGPA* and the mean attendance rate, 82, we would replace $priGPA^2$ with $(priGPA - 2.59)^2$ and $priGPA{\cdot}atndrte$ with $priGPA{\cdot}(atndrte - 82)$. The coefficient on *priGPA* becomes the partial effect at the mean values, and we would have its standard error. (See Computer Exercise C7.)

6-2d Computing Average Partial Effects

The hallmark of models with quadratics, interactions, and other nonlinear functional forms is that the partial effects depend on the values of one or more explanatory variables. For example, we just saw in Example 6.3 that the effect of *atndrte* depends on the value of *priGPA*. It is easy to see that the partial effect of *priGPA* in equation (6.18) is

$$\beta_2 + 2\beta_4 priGPA + \beta_6 atndrte$$

(something that can be verified with simple calculus or just by combining the quadratic and interaction formulas). The embellishments in equation (6.18) can be useful for seeing how the strength of associations between *stndfnl* and each explanatory variable changes with the values of all explanatory variables. The flexibility afforded by a model such as (6.18) does have a cost: it is tricky to describe the partial effects of the explanatory variables on *stndfnl* with a single number.

Often, one wants a single value to describe the relationship between the dependent variable y and each explanatory variable. One popular summary measure is the **average partial effect (APE)**, also called the *average marginal effect*. The idea behind the APE is simple for models such as (6.18). After computing the partial effect and plugging in the estimated parameters, we average the partial effects for each unit across the sample. So, the estimated partial effect of *atndrte* on *stndfnl* is

$$\hat{\beta}_1 + \hat{\beta}_6 priGPA_i.$$

We do not want to report this partial effect for each of the 680 students in our sample. Instead, we average these partial effects to obtain

$$\text{APE}_{stndfnl} = \hat{\beta}_1 + \hat{\beta}_6\overline{priGPA},$$

where $\overline{priGPA}$ is the sample average of *priGPA*. The single number $\text{APE}_{stndfnl}$ is the (estimated) APE. The APE of *priGPA* is only a little more complicated:

$$\text{APE}_{priGPA} = \hat{\beta}_2 + 2\hat{\beta}_4\overline{priGPA} + \beta_6\overline{atndrte}.$$

Both $\text{APE}_{stndfnl}$ and APE_{priGPA} tell us the size of the partial effects on average.

The centering of explanatory variables about their sample averages before creating quadratics or interactions forces the coefficient on the levels to be the APEs. This can be cumbersome in complicated models. Fortunately, some commonly used regression packages compute APEs with a simple command after OLS estimation. Just as importantly, proper standard errors are computed using the fact that an APE is a linear combination of the OLS coefficients. For example, the APEs and their standard errors for models with both quadratics and interactions, as in Example 6.3, are easy to obtain.

APEs are also useful in models that are inherently nonlinear in parameters, which we treat in Chapter 17. At that point we will revisit the definition and calculation of APEs.

6-3 More on Goodness-of-Fit and Selection of Regressors

Until now, we have not focused much on the size of R^2 in evaluating our regression models, primarily because beginning students tend to put too much weight on *R*-squared. As we will see shortly, choosing a set of explanatory variables based on the size of the *R*-squared can lead to nonsensical models. In Chapter 10, we will discover that *R*-squareds obtained from time series regressions can be artificially high and can result in misleading conclusions.

Nothing about the classical linear model assumptions requires that R^2 be above any particular value; R^2 is simply an estimate of how much variation in y is explained by $x_1, x_2, \ldots, x_k$ in the population. We have seen several regressions that have had pretty small *R*-squareds. Although this means that we have not accounted for several factors that affect y, this does not mean that the factors in u are correlated with the independent variables. The zero conditional mean assumption MLR.4 is what determines whether we get unbiased estimators of the ceteris paribus effects of the independent variables, and the size of the *R*-squared has no direct bearing on this.

A small *R*-squared does imply that the error variance is large relative to the variance of y, which means we may have a hard time precisely estimating the β_j. But remember, we saw in Section 3.4 that a large error variance can be offset by a large sample size: if we have enough data, we may be able to precisely estimate the partial effects even though we have not controlled for many unobserved factors. Whether or not we can get precise enough estimates depends on the application. For example, suppose that some incoming students at a large university are randomly given grants to buy computer equipment. If the amount of the grant is truly randomly determined, we can estimate the ceteris paribus effect of the grant amount on subsequent college grade point average by using simple regression analysis. (Because of random assignment, all of the other factors that affect GPA would be uncorrelated with the amount of the grant.) It seems likely that the grant amount would explain little of the variation in GPA, so the *R*-squared from such a regression would probably be very small. But, if we have a large sample size, we still might get a reasonably precise estimate of the effect of the grant.

Another good illustration of where poor explanatory power has nothing to do with unbiased estimation of the β_j is given by analyzing the data set APPLE. Unlike the other data sets we have used, the key explanatory variables in APPLE were set experimentally—that is, without regard to other factors that might affect the dependent variable. The variable we would like to explain, *ecolbs*, is the (hypothetical) pounds of "ecologically friendly" ("ecolabeled") apples a family would demand. Each

family (actually, family head) was presented with a description of ecolabeled apples, along with prices of regular apples (*regprc*) and prices of the hypothetical ecolabeled apples (*ecoprc*). Because the price pairs were randomly assigned to each family, they are unrelated to other observed factors (such as family income) and unobserved factors (such as desire for a clean environment). Therefore, the regression of *ecolbs* on *ecoprc*, *regprc* (across all samples generated in this way) produces unbiased estimators of the price effects. Nevertheless, the *R*-squared from the regression is only .0364: the price variables explain only about 3.6% of the total variation in *ecolbs*. So, here is a case where we explain very little of the variation in y, yet we are in the rare situation of knowing that the data have been generated so that unbiased estimation of the β_j is possible. (Incidentally, adding observed family characteristics has a very small effect on explanatory power. See Computer Exercise C11.)

Remember, though, that the relative *change* in the *R*-squared when variables are added to an equation is very useful: the *F* statistic in (4.41) for testing the joint significance crucially depends on the difference in *R*-squareds between the unrestricted and restricted models.

As we will see in Section 6.4, an important consequence of a low *R*-squared is that prediction is difficult. Because most of the variation in y is explained by unobserved factors (or at least factors we do not include in our model), we will generally have a hard time using the OLS equation to predict individual future outcomes on y given a set of values for the explanatory variables. In fact, the low *R*-squared means that we would have a hard time predicting y even if we knew the β_j, the population coefficients. Fundamentally, most of the factors that explain y are unaccounted for in the explanatory variables, making prediction difficult.

6-3a Adjusted *R*-Squared

Most regression packages will report, along with the *R*-squared, a statistic called the **adjusted *R*-squared**. Because the adjusted *R*-squared is reported in much applied work, and because it has some useful features, we cover it in this subsection.

To see how the usual *R*-squared might be adjusted, it is usefully written as

$$R^2 = 1 - (\text{SSR}/n)/(\text{SST}/n), \tag{6.20}$$

where SSR is the sum of squared residuals and SST is the total sum of squares; compared with equation (3.28), all we have done is divide both SSR and SST by n. This expression reveals what R^2 is actually estimating. Define σ_y^2 as the population variance of y and let σ_u^2 denote the population variance of the error term, u. (Until now, we have used σ^2 to denote σ_u^2, but it is helpful to be more specific here.) The **population *R*-squared** is defined as $\rho^2 = 1 - \sigma_u^2/\sigma_y^2$; this is the proportion of the variation in y in the population explained by the independent variables. This is what R^2 is supposed to be estimating.

R^2 estimates σ_u^2 by SSR/n, which we know to be biased. So why not replace SSR/n with SSR/$(n - k - 1)$? Also, we can use SST/$(n - 1)$ in place of SST/n, as the former is the unbiased estimator of σ_y^2. Using these estimators, we arrive at the adjusted *R*-squared:

$$\begin{aligned} \bar{R}^2 &= 1 - [\text{SSR}/(n - k - 1)]/[\text{SST}/(n - 1)] \\ &= 1 - \hat{\sigma}^2/[\text{SST}/(n - 1)], \end{aligned} \tag{6.21}$$

because $\hat{\sigma}^2 = \text{SSR}/(n - k - 1)$. Because of the notation used to denote the adjusted *R*-squared, it is sometimes called *R-bar squared*.

The adjusted *R*-squared is sometimes called the *corrected R-squared*, but this is not a good name because it implies that $\bar{R}^2$ is somehow better than R^2 as an estimator of the population *R*-squared. Unfortunately, $\bar{R}^2$ is *not* generally known to be a better estimator. It is tempting to think that $\bar{R}^2$ corrects the bias in R^2 for estimating the population *R*-squared, ρ^2, but it does not: the ratio of two unbiased estimators is not an unbiased estimator.

The primary attractiveness of $\bar{R}^2$ is that it imposes a penalty for adding additional independent variables to a model. We know that R^2 can never fall when a new independent variable is added to a regression equation: this is because SSR never goes up (and usually falls) as more independent variables are added (assuming we use the same set of observations). But the formula for $\bar{R}^2$ shows that it depends explicitly on k, the number of independent variables. If an independent variable is added to a regression, SSR falls, but so does the df in the regression, $n - k - 1$. SSR/$(n - k - 1)$ can go up or down when a new independent variable is added to a regression.

An interesting algebraic fact is the following: if we add a new independent variable to a regression equation, $\bar{R}^2$ increases if, and only if, the t statistic on the new variable is greater than one in absolute value. (An extension of this is that $\bar{R}^2$ increases when a group of variables is added to a regression if, and only if, the F statistic for joint significance of the new variables is greater than unity.) Thus, we see immediately that using $\bar{R}^2$ to decide whether a certain independent variable (or set of variables) belongs in a model gives us a different answer than standard t or F testing (because a t or F statistic of unity is not statistically significant at traditional significance levels).

It is sometimes useful to have a formula for $\bar{R}^2$ in terms of R^2. Simple algebra gives

$$\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1). \qquad [6.22]$$

For example, if $R^2 = .30$, $n = 51$, and $k = 10$, then $\bar{R}^2 = 1 - .70(50)/40 = .125$. Thus, for small n and large k, $\bar{R}^2$ can be substantially below R^2. In fact, if the usual R-squared is small, and $n - k - 1$ is small, $\bar{R}^2$ can actually be negative! For example, you can plug in $R^2 = .10$, $n = 51$, and $k = 10$ to verify that $\bar{R}^2 = -.125$. A negative $\bar{R}^2$ indicates a very poor model fit relative to the number of degrees of freedom.

The adjusted R-squared is sometimes reported along with the usual R-squared in regressions, and sometimes $\bar{R}^2$ is reported in place of R^2. It is important to remember that it is R^2, not $\bar{R}^2$, that appears in the F statistic in (4.41). The same formula with $\bar{R}^2_r$ and $\bar{R}^2_{ur}$ is *not* valid.

6-3b Using Adjusted *R*-Squared to Choose between Nonnested Models

In Section 4-5, we learned how to compute an F statistic for testing the joint significance of a group of variables; this allows us to decide, at a particular significance level, whether at least one variable in the group affects the dependent variable. This test does not allow us to decide *which* of the variables has an effect. In some cases, we want to choose a model without redundant independent variables, and the adjusted R-squared can help with this.

In the major league baseball salary example in Section 4-5, we saw that neither *hrunsyr* nor *rbisyr* was individually significant. These two variables are highly correlated, so we might want to choose between the models

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + u$$

and

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 rbisyr + u.$$

These two equations are **nonnested models** because neither equation is a special case of the other. The F statistics we studied in Chapter 4 only allow us to test *nested* models: one model (the restricted model) is a special case of the other model (the unrestricted model). See equations (4.32) and (4.28) for examples of restricted and unrestricted models. One possibility is to create a composite model that contains *all* explanatory variables from the original models and then to test each model against the general model using the F test. The problem with this process is that either both models might

be rejected or neither model might be rejected (as happens with the major league baseball salary example in Section 4-5). Thus, it does not always provide a way to distinguish between models with nonnested regressors.

In the baseball player salary regression using the data in MLB1, $\overline{R}^2$ for the regression containing *hrunsyr* is .6211, and $\overline{R}^2$ for the regression containing *rbisyr* is .6226. Thus, based on the adjusted *R*-squared, there is a very slight preference for the model with *rbisyr*. But the difference is practically very small, and we might obtain a different answer by controlling for some of the variables in Computer Exercise C5 in Chapter 4. (Because both nonnested models contain five parameters, the usual *R*-squared can be used to draw the same conclusion.)

Comparing $\overline{R}^2$ to choose among different nonnested sets of independent variables can be valuable when these variables represent different functional forms. Consider two models relating R&D intensity to firm sales:

$$rdintens = \beta_0 + \beta_1 \log(sales) + u. \qquad [6.23]$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u. \qquad [6.24]$$

The first model captures a diminishing return by including *sales* in logarithmic form; the second model does this by using a quadratic. Thus, the second model contains one more parameter than the first.

When equation (6.23) is estimated using the 32 observations on chemical firms in RDCHEM, R^2 is .061, and R^2 for equation (6.24) is .148. Therefore, it appears that the quadratic fits much better. But a comparison of the usual *R*-squareds is unfair to the first model because it contains one fewer parameter than (6.24). That is, (6.23) is a more parsimonious model than (6.24).

Everything else being equal, simpler models are better. Because the usual *R*-squared does not penalize more complicated models, it is better to use $\overline{R}^2$. The $\overline{R}^2$ for (6.23) is .030, while $\overline{R}^2$ for (6.24) is .090. Thus, even after adjusting for the difference in degrees of freedom, the quadratic model wins out. The quadratic model is also preferred when profit margin is added to each regression.

There is an important limitation in using $\overline{R}^2$ to choose between nonnested models: we cannot use it to choose between different functional forms for the dependent variable. This is unfortunate, because we often want to decide on whether *y* or log(*y*) (or maybe some other transformation) should be used as the dependent variable based on goodness-of-fit. But neither R^2 nor $\overline{R}^2$ can be used for this purpose. The reason is simple: these *R*-squareds measure the explained proportion of the total variation in whatever dependent variable we are using in the regression, and different nonlinear functions of the dependent variable will have different amounts of variation to explain. For example, the total variations in *y* and log(*y*) are not the same and are often very different. Comparing the adjusted *R*-squareds from regressions with these different forms of the dependent variables does not tell us anything about which model fits better; they are fitting two separate dependent variables.

GOING FURTHER 6.4

Explain why choosing a model by maximizing $\overline{R}^2$ or minimizing $\hat{\sigma}$ (the standard error of the regression) is the same thing.

EXAMPLE 6.4 CEO Compensation and Firm Performance

Consider two estimated models relating CEO compensation to firm performance:

$$\widehat{salary} = 830.63 + .0163\, sales + 19.63\, roe \qquad [6.25]$$
$$(223.90)\ \ (.0089)\ \ \ \ (11.08)$$
$$n = 209, R^2 = .029, \overline{R}^2 = .020$$

and

$$\widehat{lsalary} = 4.36 + .275\, lsales + .0179\, roe \quad [6.26]$$
$$(0.29)\quad (.033)\qquad (.0040)$$
$$n = 209,\ R^2 = .282,\ \bar{R}^2 = .275,$$

where *roe* is the return on equity discussed in Chapter 2. For simplicity, *lsalary* and *lsales* denote the natural logs of *salary* and *sales*. We already know how to interpret these different estimated equations. But can we say that one model fits better than the other?

The *R*-squared for equation (6.25) shows that *sales* and *roe* explain only about 2.9% of the variation in CEO salary in the sample. Both *sales* and *roe* have marginal statistical significance.

Equation (6.26) shows that log(*sales*) and *roe* explain about 28.2% of the variation in log(*salary*). In terms of goodness-of-fit, this much higher *R*-squared would seem to imply that model (6.26) is much better, but this is not necessarily the case. The total sum of squares for *salary* in the sample is 391,732,982, while the total sum of squares for log(*salary*) is only 66.72. Thus, there is much less variation in log(*salary*) that needs to be explained.

At this point, we can use features other than R^2 or $\bar{R}^2$ to decide between these models. For example, log(*sales*) and *roe* are much more statistically significant in (6.26) than are *sales* and *roe* in (6.25), and the coefficients in (6.26) are probably of more interest. To be sure, however, we will need to make a valid goodness-of-fit comparison.

In Section 6-4, we will offer a goodness-of-fit measure that does allow us to compare models where *y* appears in both level and log form.

6-3c Controlling for Too Many Factors in Regression Analysis

In many of the examples we have covered, and certainly in our discussion of omitted variables bias in Chapter 3, we have worried about omitting important factors from a model that might be correlated with the independent variables. It is also possible to control for too *many* variables in a regression analysis.

If we overemphasize goodness-of-fit, we open ourselves to controlling for factors in a regression model that should not be controlled for. To avoid this mistake, we need to remember the ceteris paribus interpretation of multiple regression models.

To illustrate this issue, suppose we are doing a study to assess the impact of state beer taxes on traffic fatalities. The idea is that a higher tax on beer will reduce alcohol consumption, and likewise drunk driving, resulting in fewer traffic fatalities. To measure the ceteris paribus effect of taxes on fatalities, we can model *fatalities* as a function of several factors, including the beer *tax*:

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 miles + \beta_3 percmale + \beta_4 perc16\_21 + \cdots,$$

where

miles = total miles driven.
percmale = percentage of the state population that is male.
perc16_21 = percentage of the population between ages 16 and 21, and so on.

Notice how we have not included a variable measuring per capita beer consumption. Are we committing an omitted variables error? The answer is no. If we control for beer consumption in this equation, then how would beer taxes affect traffic fatalities? In the equation

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 beercons + \cdots,$$

β_1 measures the difference in fatalities due to a one percentage point increase in *tax*, holding *beercons* fixed. It is difficult to understand why this would be interesting. We should not be controlling for differences in *beercons* across states, unless we want to test for some sort of indirect effect of beer taxes. Other factors, such as gender and age distribution, should be controlled for.

As a second example, suppose that, for a developing country, we want to estimate the effects of pesticide usage among farmers on family health expenditures. In addition to pesticide usage amounts, should we include the number of doctor visits as an explanatory variable? No. Health expenditures include doctor visits, and we would like to pick up all effects of pesticide use on health expenditures. If we include the number of doctor visits as an explanatory variable, then we are only measuring the effects of pesticide use on health expenditures other than doctor visits. It makes more sense to use number of doctor visits as a dependent variable in a separate regression on pesticide amounts.

The previous examples are what can be called **over controlling** for factors in multiple regression. Often this results from nervousness about potential biases that might arise by leaving out an important explanatory variable. But it is important to remember the ceteris paribus nature of multiple regression. In some cases, it makes no sense to hold some factors fixed precisely because they should be allowed to change when a policy variable changes.

Unfortunately, the issue of whether or not to control for certain factors is not always clear-cut. For example, Betts (1995) studies the effect of high school quality on subsequent earnings. He points out that, if better school quality results in more education, then controlling for education in the regression along with measures of quality will underestimate the return to quality. Betts does the analysis with and without years of education in the equation to get a range of estimated effects for quality of schooling.

To see explicitly how pursuing high R-squareds can lead to trouble, consider the housing price example from Section 4-5 that illustrates the testing of multiple hypotheses. In that case, we wanted to test the rationality of housing price assessments. We regressed log(*price*) on log(*assess*), log(*lotsize*), log(*sqrft*), and *bdrms* and tested whether the latter three variables had zero population coefficients while log(*assess*) had a coefficient of unity. But what if we change the purpose of the analysis and estimate a *hedonic price model*, which allows us to obtain the marginal values of various housing attributes? Should we include log(*assess*) in the equation? The adjusted R-squared from the regression with log(*assess*) is .762, while the adjusted R-squared without it is .630. Based on goodness-of-fit only, we should include log(*assess*). But this is incorrect if our goal is to determine the effects of lot size, square footage, and number of bedrooms on housing values. Including log(*assess*) in the equation amounts to holding one measure of value fixed and then asking how much an additional bedroom would change another measure of value. This makes no sense for valuing housing attributes.

If we remember that different models serve different purposes, and we focus on the ceteris paribus interpretation of regression, then we will not include the wrong factors in a regression model.

6-3d Adding Regressors to Reduce the Error Variance

We have just seen some examples of where certain independent variables should not be included in a regression model, even though they are correlated with the dependent variable. From Chapter 3, we know that adding a new independent variable to a regression can exacerbate the multicollinearity problem. On the other hand, because we are taking something out of the error term, adding a variable generally reduces the error variance. Generally, we cannot know which effect will dominate.

However, there is one case that is clear: we should always include independent variables that affect y and are *uncorrelated* with all of the independent variables of interest. Why? Because adding such a variable does not induce multicollinearity in the population (and therefore multicollinearity in the sample should be negligible), but it will reduce the error variance. In large sample sizes, the standard errors of all OLS estimators will be reduced.

As an example, consider estimating the individual demand for beer as a function of the average county beer price. It may be reasonable to assume that individual characteristics are uncorrelated with

county-level prices, and so a simple regression of beer consumption on county price would suffice for estimating the effect of price on individual demand. But it is possible to get a more precise estimate of the price elasticity of beer demand by including individual characteristics, such as age and amount of education. If these factors affect demand and are uncorrelated with price, then the standard error of the price coefficient will be smaller, at least in large samples.

As a second example, consider the grants for computer equipment given at the beginning of Section 6-3. If, in addition to the grant variable, we control for other factors that can explain college GPA, we can probably get a more precise estimate of the effect of the grant. Measures of high school grade point average and rank, SAT and ACT scores, and family background variables are good candidates. Because the grant amounts are randomly assigned, all additional control variables are uncorrelated with the grant amount; in the sample, multicollinearity between the grant amount and other independent variables should be minimal. But adding the extra controls might significantly reduce the error variance, leading to a more precise estimate of the grant effect. Remember, the issue is not unbiasedness here: we obtain an unbiased and consistent estimator whether or not we add the high school performance and family background variables. The issue is getting an estimator with a smaller sampling variance.

A related point is that when we have random assignment of a policy, we need not worry about whether some of our explanatory variables are "endogenous"—provided these variables themselves are not affected by the policy. For example, in studying the effect of hours in a job training program on labor earnings, we can include the amount of education reported prior to the job training program. We need not worry that schooling might be correlated with omitted factors, such as "ability," because we are not trying to estimate the return to schooling. We are trying to estimate the effect of the job training program, and we can include any controls that are not themselves affected by job training without biasing the job training effect. What we must avoid is including a variable such as the amount of education *after* the job training program, as some people may decide to get more education because of how many hours they were assigned to the job training program.

Unfortunately, cases where we have information on additional explanatory variables that are uncorrelated with the explanatory variables of interest are somewhat rare in the social sciences. But it is worth remembering that when these variables are available, they can be included in a model to reduce the error variance without inducing multicollinearity.

6-4 Prediction and Residual Analysis

In Chapter 3, we defined the OLS predicted or fitted values and the OLS residuals. **Predictions** are certainly useful, but they are subject to sampling variation, because they are obtained using the OLS estimators. Thus, in this section, we show how to obtain confidence intervals for a prediction from the OLS regression line.

From Chapters 3 and 4, we know that the residuals are used to obtain the sum of squared residuals and the R-squared, so they are important for goodness-of-fit and testing. Sometimes, economists study the residuals for particular observations to learn about individuals (or firms, houses, etc.) in the sample.

6.4a Confidence Intervals for Predictions

Suppose we have estimated the equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k. \qquad [6.27]$$

When we plug in particular values of the independent variables, we obtain a prediction for y, which is an estimate of the *expected value* of y given the particular values for the explanatory variables. For emphasis, let $c_1, c_2, \ldots, c_k$ denote particular values for each of the k independent variables; these

may or may not correspond to an actual data point in our sample. The parameter we would like to estimate is

$$\begin{aligned}\theta_0 &= \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k \\ &= \mathrm{E}(y|x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k).\end{aligned} \qquad [6.28]$$

The estimator of θ_0 is

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \beta_k c_k. \qquad [6.29]$$

In practice, this is easy to compute. But what if we want some measure of the uncertainty in this predicted value? It is natural to construct a confidence interval for θ_0, which is centered at $\hat{\theta}_0$.

To obtain a confidence interval for θ_0, we need a standard error for $\hat{\theta}_0$. Then, with a large *df*, we can construct a 95% confidence interval using the rule of thumb $\hat{\theta}_0 \pm 2 \cdot \mathrm{se}(\hat{\theta}_0)$. (As always, we can use the exact percentiles in a *t* distribution.)

How do we obtain the standard error of $\hat{\theta}_0$? This is the same problem we encountered in Section 4-4: we need to obtain a standard error for a linear combination of the OLS estimators. Here, the problem is even more complicated, because all of the OLS estimators generally appear in $\hat{\theta}_0$ (unless some c_j are zero). Nevertheless, the same trick that we used in Section 4-4 will work here. Write $\beta_0 = \theta_0 - \beta_1 c_1 - \cdots - \beta_k c_k$ and plug this into the equation

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

to obtain

$$y = \theta_0 + \beta_1 (x_1 - c_1) + \beta_2 (x_2 - c_2) + \cdots + \beta_k (x_k - c_k) + u. \qquad [6.30]$$

In other words, we subtract the value c_j from each observation on x_j, and then we run the regression of

$$y_i \text{ on } (x_{i1} - c_1), \ldots, (x_{ik} - c_k), i = 1, 2, \ldots, n. \qquad [6.31]$$

The predicted value in (6.29) and, more importantly, its standard error, are obtained from the *intercept* (or constant) in regression (6.31).

As an example, we obtain a confidence interval for a prediction from a college GPA regression, where we use high school information.

EXAMPLE 6.5 Confidence Interval for Predicted College GPA

Using the data in GPA2, we obtain the following equation for predicting college GPA:

$$\begin{aligned}\widehat{colgpa} &= 1.493 + .00149\, sat - .01386\, hsperc \\ &\qquad (0.075)\ \ (.00007)\ \ \ \ (.00056) \\ &\quad - .06088\, hsize + .00546\, hsize^2 \\ &\qquad (.01650)\ \ \ \ \ (.00227) \\ n &= 4{,}137,\ R^2 = .278,\ \overline{R}^2 = .277,\ \hat{\sigma} = .560,\end{aligned} \qquad [6.32]$$

where we have reported estimates to several digits to reduce round-off error. What is predicted college GPA, when $sat = 1{,}200$, $hsperc = 30$, and $hsize = 5$ (which means 500)? This is easy to get by plugging these values into equation (6.32): $\widehat{colgpa} = 2.70$ (rounded to two digits). Unfortunately, we cannot use equation (6.32) directly to get a confidence interval for the expected *colgpa* at the given

values of the independent variables. One simple way to obtain a confidence interval is to define a new set of independent variables: $sat0 = sat - 1{,}200$, $hsperc0 = hsperc - 30$, $hsize0 = hsize - 5$, and $hsizesq0 = hsize^2 - 25$. When we regress *colgpa* on these new independent variables, we get

$$\widehat{colgpa} = 2.700 + .00149\, sat0 - .01386\, hsperc0$$
$$(0.020)\ (.00007)\qquad (.00056)$$
$$- .06088\, hsize0 + .00546\, hsizesq0$$
$$(.01650)\qquad (.00227)$$
$$n = 4{,}137,\ R^2 = .278,\ \overline{R}^2 = .277,\ \hat{\sigma} = .560.$$

The only difference between this regression and that in (6.32) is the intercept, which is the prediction we want, along with its standard error, .020. It is not an accident that the slope coefficients, their standard errors, *R*-squared, and so on are the same as before; this provides a way to check that the proper transformations were done. We can easily construct a 95% confidence interval for the expected college GPA: $2.70 \pm 1.96(.020)$ or about 2.66 to 2.74. This confidence interval is rather narrow due to the very large sample size.

Because the variance of the intercept estimator is smallest when each explanatory variable has zero sample mean (see Problem 10, part (iv) in Chapter 2 for the simple regression case), it follows from the regression in (6.31) that the variance of the prediction is smallest at the mean values of the x_j. (That is, $c_j = \bar{x}_j$ for all j.) This result is not too surprising, as we have the most faith in our regression line near the middle of the data. As the values of the c_j get farther away from the $\bar{x}_j$, $\text{Var}(\hat{y})$ gets larger and larger.

The previous method allows us to put a confidence interval around the OLS estimate of $\text{E}(y|x_1, \ldots, x_k)$ for any values of the explanatory variables. In other words, we obtain a confidence interval for the *average* value of y for the subpopulation with a given set of covariates. But a confidence interval for the average person in the subpopulation is not the same as a confidence interval for a particular unit (individual, family, firm, and so on) from the population. In forming a confidence interval for an unknown outcome on y, we must account for another very important source of variation: the variance in the unobserved error, which measures our ignorance of the unobserved factors that affect y.

Let y^0 denote the value for which we would like to construct a confidence interval, which we sometimes call a **prediction interval**. For example, y^0 could represent a person or firm not in our original sample. Let $x_1^0, \ldots, x_k^0$ be the new values of the independent variables, which we assume we observe, and let u^0 be the unobserved error. Therefore, we have

$$y^0 = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \cdots + \beta_k x_k^0 + u^0. \qquad [6.33]$$

As before, our best prediction of y^0 is the expected value of y^0 given the explanatory variables, which we estimate from the OLS regression line: $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \hat{\beta}_2 x_2^0 + \cdots + \hat{\beta}_k x_k^0$. The **prediction error** in using $\hat{y}^0$ to predict y^0 is

$$\hat{e}^0 = y^0 - \hat{y}^0 = (\beta_0 + \beta_1 x_1^0 + \cdots + \beta_k x_k^0) + u^0 - \hat{y}^0. \qquad [6.34]$$

Now, $\text{E}(\hat{y}^0) = \text{E}(\hat{\beta}_0) + \text{E}(\hat{\beta}_1)x_1^0 + \text{E}(\hat{\beta}_2)x_2^0 + \cdots + \text{E}(\hat{\beta}_k)x_k^0 = \beta_0 + \beta_1 x_1^0 + \cdots + \beta_k x_k^0$, because the $\hat{\beta}_j$ are unbiased. (As before, these expectations are all conditional on the sample values of the independent variables.) Because u^0 has zero mean, $\text{E}(\hat{e}^0) = 0$. We have shown that the expected prediction error is zero.

In finding the variance of $\hat{e}^0$, note that u^0 is uncorrelated with each $\hat{\beta}_j$, because u^0 is uncorrelated with the errors in the sample used to obtain the $\hat{\beta}_j$. By basic properties of covariance (see Math Refresher B),

u^0 and $\hat{y}^0$ are uncorrelated. Therefore, the **variance of the prediction error** (conditional on all in-sample values of the independent variables) is the sum of the variances:

$$\text{Var}(\hat{e}^0) = \text{Var}(\hat{y}^0) + \text{Var}(u^0) = \text{Var}(\hat{y}^0) + \sigma^2, \qquad [6.35]$$

where $\sigma^2 = \text{Var}(u^0)$ is the error variance. There are two sources of variation in $\hat{e}^0$. The first is the sampling error in $\hat{y}^0$, which arises because we have estimated the β_j. Because each $\hat{\beta}_j$ has a variance proportional to $1/n$, where n is the sample size, $\text{Var}(\hat{y}^0)$ is proportional to $1/n$. This means that, for large samples, $\text{Var}(\hat{y}^0)$ can be very small. By contrast, σ^2 is the variance of the error in the population; it does not change with the sample size. In many examples, σ^2 will be the dominant term in (6.35).

Under the classical linear model assumptions, the $\hat{\beta}_j$ and u^0 are normally distributed, and so $\hat{e}^0$ is also normally distributed (conditional on all sample values of the explanatory variables). Earlier, we described how to obtain an unbiased estimator of $\text{Var}(\hat{y}^0)$, and we obtained our unbiased estimator of σ^2 in Chapter 3. By using these estimators, we can define the standard error of $\hat{e}^0$ as

$$\text{se}(\hat{e}^0) = \{[\text{se}(\hat{y}^0)]^2 + \hat{\sigma}^2\}^{1/2}. \qquad [6.36]$$

Using the same reasoning for the t statistics of the $\hat{\beta}_j$, $\hat{e}^0/\text{se}(\hat{e}^0)$ has a t distribution with $n - (k + 1)$ degrees of freedom. Therefore,

$$\text{P}[-t_{.025} \leq \hat{e}^0/\text{se}(\hat{e}^0) \leq t_{.025}] = .95,$$

where $t_{.025}$ is the 97.5th percentile in the t_{n-k-1} distribution. For large $n - k - 1$, remember that $t_{.025} \approx 1.96$. Plugging in $\hat{e}^0 = y^0 - \hat{y}^0$ and rearranging gives a 95% prediction interval for y^0:

$$\hat{y}^0 \pm t_{.025} \cdot \text{se}(\hat{e}^0); \qquad [6.37]$$

as usual, except for small *df*, a good rule of thumb is $\hat{y}^0 \pm 2\text{se}(\hat{e}^0)$. This is wider than the confidence interval for $\hat{y}^0$ itself because of $\hat{\sigma}^2$ in (6.36); it often is much wider to reflect the factors in u^0 that we have not accounted for.

EXAMPLE 6.6 Confidence Interval for Future College GPA

Suppose we want a 95% confidence interval for the future college GPA of a high school student with *sat* = 1,200, *hsperc* = 30, and *hsize* = 5. In Example 6.5, we obtained a 95% CI for the *average* college GPA among all students with the particular characteristics *sat* = 1,200, *hsperc* = 30, and *hsize* = 5. Now, we want a 95% CI for any *particular* student with these characteristics. The 95% prediction interval must account for the variation in the individual, unobserved characteristics that affect college performance. We have everything we need to obtain a CI for *colgpa*. $\text{se}(\hat{y}^0) = .020$ and $\hat{\sigma} = .560$ and so, from (6.36), $\text{se}(\hat{e}^0) = [(.020)^2 + (.560)^2]^{1/2} \approx .560$. Notice how small $\text{se}(\hat{y}^0)$ is relative to $\hat{\sigma}$: virtually all of the variation in $\hat{e}^0$ comes from the variation in u^0. The 95% CI is $2.70 \pm 1.96(.560)$ or about 1.60 to 3.80. This is a wide confidence interval and shows that, based on the factors we included in the regression, we cannot accurately pin down an individual's future college grade point average. (In one sense, this is good news, as it means that high school rank and performance on the SAT do not preordain one's performance in college.) Evidently, the unobserved characteristics that affect college GPA vary widely among individuals with the same observed SAT score and high school rank.

6-4b Residual Analysis

Sometimes, it is useful to examine individual observations to see whether the actual value of the dependent variable is above or below the predicted value; that is, to examine the residuals for the individual observations. This process is called **residual analysis**. Economists have been known to examine the residuals from a regression in order to aid in the purchase of a home. The following housing price example illustrates residual analysis. Housing price is related to various observable characteristics of the house. We can list all of the characteristics that we find important, such as size, number of bedrooms, number of bathrooms, and so on. We can use a sample of houses to estimate a relationship between price and attributes, where we end up with a predicted value and an actual value for each house. Then, we can construct the residuals, $\hat{u}_i = y_i - \hat{y}_i$. The house with the most negative residual is, at least based on the factors we have controlled for, the most underpriced one relative to its *observed* characteristics. Of course, a selling price substantially below its predicted price could indicate some undesirable feature of the house that we have failed to account for, and which is therefore contained in the unobserved error. In addition to obtaining the prediction and residual, it also makes sense to compute a confidence interval for what the future selling price of the home could be, using the method described in equation (6.37).

Using the data in HPRICE1, we run a regression of *price* on *lotsize*, *sqrft*, and *bdrms*. In the sample of 88 homes, the most negative residual is −120.206, for the 81st house. Therefore, the asking price for this house is $120,206 below its predicted price.

There are many other uses of residual analysis. One way to rank law schools is to regress median starting salary on a variety of student characteristics (such as median LSAT scores of entering class, median college GPA of entering class, and so on) and to obtain a predicted value and residual for each law school. The law school with the largest residual has the highest predicted value added. (Of course, there is still much uncertainty about how an individual's starting salary would compare with the median for a law school overall.) These residuals can be used along with the costs of attending each law school to determine the best value; this would require an appropriate discounting of future earnings.

Residual analysis also plays a role in legal decisions. A *New York Times* article entitled "Judge Says Pupil's Poverty, Not Segregation, Hurts Scores" (6/28/95) describes an important legal case. The issue was whether the poor performance on standardized tests in the Hartford School District, relative to performance in surrounding suburbs, was due to poor school quality at the highly segregated schools. The judge concluded that "the disparity in test scores does not indicate that Hartford is doing an inadequate or poor job in educating its students or that its schools are failing, because the predicted scores based upon the relevant socioeconomic factors are about at the levels that one would expect." This conclusion is based on a regression analysis of average or median scores on socioeconomic characteristics of various school districts in Connecticut. The judge's conclusion suggests that, given the poverty levels of students at Hartford schools, the actual test scores were similar to those predicted from a regression analysis: the residual for Hartford was not sufficiently negative to conclude that the schools themselves were the cause of low test scores.

GOING FURTHER 6.5

How would you use residual analysis to determine which professional athletes are overpaid or underpaid relative to their performance?

6-4c Predicting *y* When log(*y*) Is the Dependent Variable

Because the natural log transformation is used so often for the dependent variable in empirical economics, we devote this subsection to the issue of predicting y when $\log(y)$ is the dependent variable. As a byproduct, we will obtain a goodness-of-fit measure for the log model that can be compared with the R-squared from the level model.

To obtain a prediction, it is useful to define $logy = \log(y)$; this emphasizes that it is the log of y that is predicted in the model

$$logy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u. \qquad [6.38]$$

In this equation, the x_j might be transformations of other variables; for example, we could have $x_1 = \log(sales)$, $x_2 = \log(mktval)$, $x_3 = ceoten$ in the CEO salary example.

Given the OLS estimators, we know how to predict *logy* for any value of the independent variables:

$$\widehat{logy} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k. \qquad [6.39]$$

Now, because the exponential undoes the log, our first guess for predicting y is to simply exponentiate the predicted value for $\log(y)$: $\hat{y} = \exp(\widehat{logy})$. This does not work; in fact, it will systematically *underestimate* the expected value of y. In fact, if model (6.38) follows the CLM assumptions MLR.1 through MLR.6, it can be shown that

$$\mathrm{E}(y|\boldsymbol{x}) = \exp(\sigma^2/2)\cdot\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k),$$

where $\boldsymbol{x}$ denotes the independent variables and σ^2 is the variance of u. [*If* $u \sim \text{Normal}(0,\sigma^2)$, then the expected value of $\exp(u)$ is $\exp(\sigma^2/2)$.] This equation shows that a simple adjustment is needed to predict y:

$$\hat{y} = \exp(\hat{\sigma}^2/2)\exp(\widehat{logy}), \qquad [6.40]$$

where $\hat{\sigma}^2$ is simply the unbiased estimator of σ^2. Because $\hat{\sigma}$, the standard error of the regression, is always reported, obtaining predicted values for y is easy. Because $\hat{\sigma}^2 > 0$, $\exp(\hat{\sigma}^2/2) > 1$. For large $\hat{\sigma}^2$, this adjustment factor can be substantially larger than unity.

The prediction in (6.40) is not unbiased, but it is consistent. There are no unbiased predictions of y, and in many cases, (6.40) works well. However, it does rely on the normality of the error term, u. In Chapter 5, we showed that OLS has desirable properties, even when u is not normally distributed. Therefore, it is useful to have a prediction that does not rely on normality. If we just assume that u is independent of the explanatory variables, then we have

$$\mathrm{E}(y|\boldsymbol{x}) = \alpha_0\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k), \qquad [6.41]$$

where α_0 is the expected value of $\exp(u)$, which must be greater than unity.

Given an estimate $\hat{\alpha}_0$, we can predict y as

$$\hat{y} = \hat{\alpha}_0\exp(\widehat{logy}), \qquad [6.42]$$

which again simply requires exponentiating the predicted value from the log model and multiplying the result by $\hat{\alpha}_0$.

Two approaches suggest themselves for estimating α_0 without the normality assumption. The first is based on $\alpha_0 = \mathrm{E}[\exp(u)]$. To estimate α_0 we replace the population expectation with a sample average and then we replace the unobserved errors, u_i, with the OLS residuals, $\hat{u}_i = \log(y_i) - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}$. This leads to the method of moments estimator (see Math Refresher C)

$$\hat{\alpha}_0 = n^{-1}\sum_{i=1}^{n}\exp(\hat{u}_i). \qquad [6.43]$$

Not surprisingly, $\hat{\alpha}_0$ is a consistent estimator of α_0, but it is not unbiased because we have replaced u_i with $\hat{u}_i$ inside a nonlinear function. This version of $\hat{\alpha}_0$ is a special case of what Duan (1983) called a **smearing estimate**. Because the OLS residuals have a zero sample average, it can be shown that, for any data set, $\hat{\alpha}_0 > 1$. (Technically, $\hat{\alpha}_0$ would equal one if all the OLS residuals were zero, but this

never happens in any interesting application.) That $\hat{\alpha}_0$ is necessarily greater than one is convenient because it must be that $\alpha_0 > 1$.

A different estimate of α_0 is based on a simple regression through the origin. To see how it works, define $m_i = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})$, so that, from equation (6.41), $\mathrm{E}(y_i|m_i) = \alpha_0 m_i$. If we could observe the m_i, we could obtain an unbiased estimator of α_0 from the regression y_i on m_i without an intercept. Instead, we replace the β_j with their OLS estimates and obtain $\hat{m}_i = \exp(\widehat{logy}_i)$, where, of course, the $\widehat{logy}_i$ are the fitted values from the regression $logy_i$ on $x_{i1}, \ldots, x_{ik}$ (with an intercept). Then $\check{\alpha}_0$ [to distinguish it from $\hat{\alpha}_0$ in equation (6.43)] is the OLS slope estimate from the simple regression y_i on $\hat{m}_i$ (no intercept):

$$\check{\alpha}_0 = \left(\sum_{i=1}^{n} \hat{m}_i^2\right)^{-1}\left(\sum_{i=1}^{n} \hat{m}_i y_i\right). \qquad [6.44]$$

We will call $\check{\alpha}_0$ the regression estimate of α_0. Like $\hat{\alpha}_0$, $\check{\alpha}_0$ is consistent but not unbiased. Interestingly, $\check{\alpha}_0$ is not guaranteed to be greater than one, although it will be in most applications. If $\check{\alpha}_0$ is less than one, and especially if it is much less than one, it is likely that the assumption of independence between u and the x_j is violated. If $\check{\alpha}_0 < 1$, one possibility is to just use the estimate in (6.43), although this may simply be masking a problem with the linear model for log(y).

We summarize the steps:

6-4d Predicting *y* When the Dependent Variable Is log(*y*)

1. Obtain the fitted values, $\widehat{logy}_i$, and residuals, $\hat{u}_i$, from the regression $logy$ on $x_1, \ldots, x_k$.
2. Obtain $\hat{\alpha}_0$ as in equation (6.43) or $\check{\alpha}_0$ in equation (6.44).
3. For given values of $x_1, \ldots, x_k$, obtain $\widehat{logy}$ from (6.42).
4. Obtain the prediction $\hat{y}$ from (6.42) (with $\hat{\alpha}_0$ or $\check{\alpha}_0$).

We now show how to predict CEO salaries using this procedure.

EXAMPLE 6.7 Predicting CEO Salaries

The model of interest is

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + \beta_3 ceoten + u,$$

so that β_1 and β_2 are elasticities and $100 \cdot \beta_3$ is a semi-elasticity. The estimated equation using CEOSAL2 is

$$\begin{aligned} \widehat{lsalary} &= 4.504 + .163\ lsales + .109\ lmktval + .0117\ ceoten \\ &\qquad (.257)\ \ (.039)\qquad\ (.050)\qquad\ (.0053) \qquad [6.45] \\ n &= 177, R^2 = .318, \end{aligned}$$

where, for clarity, we let *lsalary* denote the log of *salary*, and similarly for *lsales* and *lmktval*. Next, we obtain $\hat{m}_i = \exp(\widehat{lsalary}_i)$ for each observation in the sample.

The Duan smearing estimate from (6.43) is about $\hat{\alpha}_0 = 1.136$, and the regression estimate from (6.44) is $\check{\alpha}_0 = 1.117$. We can use either estimate to predict *salary* for any values of *sales*, *mktval*, and *ceoten*. Let us find the prediction for *sales* = 5,000 (which means \$5 billion because *sales* is in millions), *mktval* = 10,000 (or \$10 billion), and *ceoten* = 10. From (6.45), the prediction for *lsalary* is $4.504 + .163 \cdot \log(5{,}000) + .109 \cdot \log(10.000) + .0117(10) \approx 7.013$, and $\exp(7.013) \approx 1{,}110.983$. Using the estimate of α_0 from (6.43), the predicted salary is about 1,262.077, or \$1,262,077. Using the estimate from (6.44) gives an estimated salary of about \$1,240,968. These differ from each other by much less than each differs from the naive prediction of \$1,110,983.

We can use the previous method of obtaining predictions to determine how well the model with $\log(y)$ as the dependent variable explains y. We already have measures for models when y is the dependent variable: the R-squared and the adjusted R-squared. The goal is to find a goodness-of-fit measure in the $\log(y)$ model that can be compared with an R-squared from a model where y is the dependent variable.

There are different ways to define a goodness-of-fit measure after retransforming a model for $\log(y)$ to predict y. Here we present two approaches that are easy to implement. The first gives the same goodness-of-fit measures whether we estimate α_0 as in (6.40), (6.43), or (6.44). To motivate the measure, recall that in the linear regression equation estimated by OLS,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k, \qquad [6.46]$$

the usual R-squared is simply the square of the correlation between y_i and $\hat{y}_i$ (see Section 3-2). Now, if instead we compute fitted values from (6.42)—that is, $\hat{y}_i = \hat{\alpha}_0 m_i$ for all observations i—then it makes sense to use the square of the correlation between y_i and these fitted values as an R-squared. Because correlation is unaffected if we multiply by a constant, it does not matter which estimate of α_0 we use. In fact, this R-squared measure for y [not $\log(y)$] is just the squared correlation between y_i and $\hat{m}_i$. We can compare this directly with the R-squared from equation (6.46).

The squared correlation measure does not depend on how we estimate α_0. A second approach is to compute an R-squared for y based on a sum of squared residuals. For concreteness, suppose we use equation (6.43) to estimate α_0. Then the residual for predicting y_i is

$$\hat{r}_i = y_i - \hat{\alpha}_0 \exp(\widehat{logy}_i), \qquad [6.47]$$

and we can use these residuals to compute a sum of squared residuals. Using the formula for R-squared from linear regression, we are led to

$$1 - \frac{\sum_{i=1}^{n} \hat{r}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})} \qquad [6.48]$$

as an alternative goodness-of-fit measure that can be compared with the R-squared from the linear model for y. Notice that we can compute such a measure for the alternative estimates of α_0 in equation (6.40) and (6.44) by inserting those estimates in place of $\hat{\alpha}_0$ in (6.47). Unlike the squared correlation between y_i and $\hat{m}_i$, the R-squared in (6.48) will depend on how we estimate α_0. The estimate that minimizes $\sum_{i=1}^{n} \hat{r}_i^2$ is that in equation (6.44), but that does not mean we should prefer it (and certainly not if $\check{\alpha}_0 < 1$). We are not really trying to choose among the different estimates of α_0; rather, we are finding goodness-of-fit measures that can be compared with the linear model for y.

EXAMPLE 6.8 Predicting CEO Salaries

After we obtain the $\hat{m}_i$, we just obtain the correlation between $salary_i$ and $\hat{m}_i$; it is .493. The square of it is about .243, and this is a measure of how well the log model explains the variation in *salary*, not $\log(salary)$. [The R^2 from (6.45), .318, tells us that the log model explains about 31.8% of the variation in $\log(salary)$.]

As a competing linear model, suppose we estimate a model with all variables in levels:

$$salary = \beta_0 + \beta_1 sales + \beta_2 mktval + \beta_3 ceoten + u. \qquad [6.49]$$

The key is that the dependent variable is *salary*. We could use logs of *sales* or *mktval* on the right-hand side, but it makes more sense to have all dollar values in levels if one (*salary*) appears as a level. The R-squared from estimating this equation using the same 177 observations is .201. Thus, the log model explains more of the variation in *salary*, and so we prefer it to (6.49) on goodness-of-fit

grounds. The log model is also preferred because it seems more realistic and its parameters are easier to interpret.

If we maintain the full set of classical linear model assumptions in the model (6.38), we can easily obtain prediction intervals for $y^0 = \exp(\beta_0 + \beta_1 x_1^0 + \cdots + \beta_k x_k^0 + u^0)$ when we have estimated a linear model for $\log(y)$. Recall that $x_1^0, x_2^0, \ldots, x_k^0$ are known values and u^0 is the unobserved error that partly determines y^0. From equation (6.37), a 95% prediction interval for $logy^0 = \log(y^0)$ is simply $\widehat{logy}^0 \pm t_{.025} \cdot \text{se}(\hat{e}^0)$, where $\text{se}(\hat{e}^0)$ is obtained from the regression of $\log(y)$ on $x_1, \ldots, x_k$ using the original n observations. Let $c_l = \widehat{logy}^0 - t_{.025} \cdot \text{se}(\hat{e}^0)$ and $c_u = \widehat{logy}^0 + t_{.025} \cdot \text{se}(\hat{e}^0)$ be the lower and upper bounds of the prediction interval for $logy^0$. That is, $\text{P}(c_l \leq logy^0 \leq c_u) = .95$. Because the exponential function is strictly increasing, it is also true that $\text{P}[\exp(c_l) \leq \exp(logy^0) \leq \exp(c_u)] = .95$, that is, $\text{P}[\exp(c_l) \leq y^0 \leq \exp(c_u)] = .95$. Therefore, we can take $\exp(c_l)$ and $\exp(c_u)$ as the lower and upper bounds, respectively, for a 95% prediction interval for y^0. For large n, $t_{.025} = 1.96$, and so a 95% prediction interval for y^0 is $\exp[-1.96 \cdot \text{se}(\hat{e}^0)]\exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$ to $\exp[-1.96 \cdot \text{se}(\hat{e}^0)]\exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$, where $\mathbf{x}^0\hat{\boldsymbol{\beta}}$ is shorthand for $\hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0$. Remember, the $\hat{\beta}_j$ and $\text{se}(\hat{e}^0)$ are obtained from the regression with $\log(y)$ as the dependent variable. Because we assume normality of u in (6.38), we probably would use (6.40) to obtain a point prediction for y^0. Unlike in equation (6.37), this point prediction will not lie halfway between the lower and upper bounds $\exp(c_l)$ and $\exp(c_u)$. One can obtain different 95% prediction intervalues by choosing different quantiles in the t_{n-k-1} distribution. If $q_{\alpha 1}$ and $q_{\alpha 2}$ are quantiles with $\alpha_2 - \alpha_1 = .95$, then we can choose $c_l = q_{\alpha 1}\text{se}(\hat{e}^0)$ and $c_u = q_{\alpha 2}\text{se}(\hat{e}^0)$.

As an example, consider the CEO salary regression, where we make the prediction at the same values of *sales*, *mktval*, and *ceoten* as in Example 6.7. The standard error of the regression for (6.43) is about .505, and the standard error of $\widehat{logy}^0$ is about .075. Therefore, using equation (6.36), $\text{se}(\hat{e}^0) \approx .511$; as in the GPA example, the error variance swamps the estimation error in the parameters, even though here the sample size is only 177. A 95% prediction interval for $salary^0$ is $\exp[-1.96 \cdot (.511)]\exp(7.013)$ to $\exp[1.96 \cdot (.511)]\exp(7.013)$, or about 408.071 to 3,024.678, that is, \$408,071 to \$3,024,678. This very wide 95% prediction interval for CEO salary at the given sales, market value, and tenure values shows that there is much else that we have not included in the regression that determines salary. Incidentally, the point prediction for salary, using (6.40), is about \$1,262,075—higher than the predictions using the other estimates of α_0 and closer to the lower bound than the upper bound of the 95% prediction interval.

Summary

In this chapter, we have covered some important multiple regression analysis topics.

Section 6-1 showed that a change in the units of measurement of an independent variable changes the OLS coefficient in the expected manner: if x_j is multiplied by c, its coefficient is divided by c. If the dependent variable is multiplied by c, *all* OLS coefficients are multiplied by c. Neither t nor F statistics are affected by changing the units of measurement of any variables.

We discussed beta coefficients, which measure the effects of the independent variables on the dependent variable in standard deviation units. The beta coefficients are obtained from a standard OLS regression after the dependent and independent variables have been transformed into z-scores.

We provided a detailed discussion of functional form, including the logarithmic transformation, quadratics, and interaction terms. It is helpful to summarize some of our conclusions.

CONSIDERATIONS WHEN USING LOGARITHMS

1. The coefficients have percentage change interpretations. We can be ignorant of the units of measurement of any variable that appears in logarithmic form, and changing units from, say, dollars to thousands of dollars has no effect on a variable's coefficient when that variable appears in logarithmic form.

2. Logs are often used for dollar amounts that are always positive, as well as for variables such as population, especially when there is a lot of variation. They are used less often for variables measured in years, such as schooling, age, and experience. Logs are used infrequently for variables that are already percents or proportions, such as an unemployment rate or a pass rate on a test.
3. Models with $\log(y)$ as the dependent variable often more closely satisfy the classical linear model assumptions. For example, the model has a better chance of being linear, homoskedasticity is more likely to hold, and normality is often more plausible.
4. In many cases, taking the log greatly reduces the variation of a variable, making OLS estimates less prone to outlier influence. However, in cases where y is a fraction and close to zero for many observations, $\log(y_i)$ can have much more variability than y_i. For values y_i very close to zero, $\log(y_i)$ is a negative number very large in magnitude.
5. If $y \geq 0$ but $y = 0$ is possible, we cannot use $\log(y)$. Sometimes $\log(1 + y)$ is used, but interpretation of the coefficients is difficult.
6. For large changes in an explanatory variable, we can compute a more accurate estimate of the percentage change effect.
7. It is harder (but possible) to predict y when we have estimated a model for $\log(y)$.

CONSIDERATIONS WHEN USING QUADRATICS

1. A quadratic function in an explanatory variable allows for an increasing or decreasing effect.
2. The turning point of a quadratic is easily calculated, and it should be calculated to see if it makes sense.
3. Quadratic functions where the coefficients have the opposite sign have a strictly positive turning point; if the signs of the coefficients are the same, the turning point is at a negative value of x.
4. A seemingly small coefficient on the square of a variable can be practically important in what it implies about a changing slope. One can use a t test to see if the quadratic is statistically significant, and compute the slope at various values of x to see if it is practically important.
5. For a model quadratic in a variable x, the coefficient on x measures the partial effect starting from $x = 0$, as can be seen in equation (6.11). If zero is not a possible or interesting value of x, one can center x about a more interesting value, such as the average in the sample, before computing the square. This is the same as computing the average partial effect. Computing Exercise C12 provides an example.

CONSIDERATIONS WHEN USING INTERACTIONS

1. Interaction terms allow the partial effect of an explanatory variable, say x_1, to depend on the level of another variable, say x_2—and vice versa.
2. Interpreting models with interactions can be tricky. The coefficient on x_1, say β_1, measures the partial effect of x_1 on y when $x_2 = 0$, which may be impossible or uninteresting. Centering x_1 and x_2 around interesting values before constructing the interaction term typically leads to an equation that is visually more appealing. When the variables are centered about their sample averages before multiplying them together to create the interaction, the coefficients on the levels become estimated average partial effects.
3. A standard t test can be used to determine if an interaction term is statistically significant. Computing the partial effects at different values of the explanatory variables can be used to determine the practical importance of interactions.

We introduced the adjusted R-squared, $\overline{R}^2$, as an alternative to the usual R-squared for measuring goodness-of-fit. Whereas R^2 can never fall when another variable is added to a regression, $\overline{R}^2$ penalizes the number of regressors and can drop when an independent variable is added. This makes $\overline{R}^2$ preferable for choosing between nonnested models with different numbers of explanatory variables. Neither R^2 nor $\overline{R}^2$ can be used to compare models with different dependent variables. Nevertheless, it is fairly easy to obtain goodness-of-fit measures for choosing between y and $\log(y)$ as the dependent variable, as shown in Section 6-4.

In Section 6-3, we discussed the somewhat subtle problem of relying too much on R^2 or $\overline{R}^2$ in arriving at a final model: it is possible to control for too many factors in a regression model. For this reason, it is important to think ahead about model specification, particularly the ceteris paribus nature of the multiple regression equation. Explanatory variables that affect y and are uncorrelated with all the other explanatory variables can be used to reduce the error variance without inducing multicollinearity.

In Section 6-4, we demonstrated how to obtain a confidence interval for a prediction made from an OLS regression line. We also showed how a confidence interval can be constructed for a future, unknown value of y.

Occasionally, we want to predict y when $\log(y)$ is used as the dependent variable in a regression model. Section 6-4 explains this simple method. Finally, we are sometimes interested in knowing about the sign and magnitude of the residuals for particular observations. Residual analysis can be used to determine whether particular members of the sample have predicted values that are well above or well below the actual outcomes.

Key Terms

Adjusted *R*-Squared
Average Partial Effect (APE)
Beta Coefficients
Bootstrap
Bootstrap Standard Error
Interaction Effect
Nonnested Models
Over Controlling
Population *R*-Squared
Prediction Error
Prediction Interval
Predictions
Quadratic Functions
Resampling Method
Residual Analysis
Smearing Estimate
Standardized Coefficients
Variance of the Prediction Error

Problems

1 The following equation was estimated using the data in CEOSAL1:

$$\widehat{\log(salary)} = \underset{(.324)}{4.322} + \underset{(.033)}{.276}\log(sales) + \underset{(.0129)}{.0215}\,roe - \underset{(.00026)}{.00008}\,roe^2$$
$$n = 209,\ R^2 = .282.$$

This equation allows *roe* to have a diminishing effect on $\log(salary)$. Is this generality necessary? Explain why or why not.

2 Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ be the OLS estimates from the regression of y_i on $x_{i1}, \ldots, x_{ik}$, $i = 1, 2, \ldots, n$. For nonzero constants $c_1, \ldots, c_k$, argue that the OLS intercept and slopes from the regression of $c_0 y_i$ on $c_1 x_{i1}, \ldots, c_k x_{ik}$, $i = 1, 2, \ldots, n$, are given by $\tilde{\beta}_0 = c_0\hat{\beta}_0, \tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \hat{\beta}_k = (c_0/c_k)\hat{\beta}_k$. [*Hint*: Use the fact that the $\hat{\beta}_j$ solve the first order conditions in (3.13), and the $\tilde{\beta}_j$ must solve the first order conditions involving the rescaled dependent and independent variables.]

3 Using the data in RDCHEM, the following equation was obtained by OLS:

$$\widehat{rdintens} = \underset{(.429)}{2.613} + \underset{(.00014)}{.00030}\,sales - \underset{(.0000000037)}{.0000000070}\,sales^2$$
$$n = 32,\ R^2 = .1484.$$

(i) At what point does the marginal effect of *sales* on *rdintens* become negative?
(ii) Would you keep the quadratic term in the model? Explain.

(iii) Define *salesbil* as sales measured in billions of dollars: $salesbil = sales/1{,}000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report standard errors and the *R*-squared. [*Hint*: Note that $salesbil^2 = sales^2/(1{,}000)^2$.]

(iv) For the purpose of reporting the results, which equation do you prefer?

4 The following model allows the return to education to depend upon the total amount of both parents' education, called *pareduc*:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ \cdot pareduc + \beta_3 exper + \beta_4 tenure + u.$$

(i) Show that, in decimal form, the return to another year of education in this model is

$$\Delta\log(wage)/\Delta educ = \beta_1 + \beta_2 pareduc.$$

What sign do you expect for β_2? Why?

(ii) Using the data in WAGE2, the estimated equation is

$$\begin{aligned}\widehat{\log(wage)} &= 5.65 + .047\,educ + .00078\,educ\cdot pareduc + \\ &\quad (.13)\ (.010)\qquad (.00021) \\ &\quad .019\,exper + .010\,tenure \\ &\quad (.004)\qquad (.003) \\ n &= 722,\ R^2 = .169.\end{aligned}$$

(Only 722 observations contain full information on parents' education.) Interpret the coefficient on the interaction term. It might help to choose two specific values for *pareduc*—for example, $pareduc = 32$ if both parents have a college education, or $pareduc = 24$ if both parents have a high school education—and to compare the estimated return to *educ*.

(iii) When *pareduc* is added as a separate variable to the equation, we get:

$$\begin{aligned}\widehat{\log(wage)} &= 4.94 + .097\,educ + .033\,pareduc - .0016\,educ\cdot pareduc \\ &\quad (.38)\ (.027)\qquad (.017)\qquad (.0012) \\ &\quad + .020\,exper + .010\,tenure \\ &\quad (.004)\qquad (.003) \\ n &= 722,\ R^2 = .174.\end{aligned}$$

Does the estimated return to education now depend positively on parent education? Test the null hypothesis that the return to education does not depend on parent education.

5 In Example 4.2, where the percentage of students receiving a passing score on a tenth-grade math exam (*math*10) is the dependent variable, does it make sense to include *sci11*—the percentage of eleventh graders passing a science exam—as an additional explanatory variable?

6 When $atndrte^2$ and $ACT\cdot atndrte$ are added to the equation estimated in (6.19), the *R*-squared becomes .232. Are these additional terms jointly significant at the 10% level? Would you include them in the model?

7 The following three equations were estimated using the 1,534 observations in 401K:

$$\begin{aligned}\widehat{prate} &= 80.29 + 5.44\,mrate + .269\,age - .00013\,totemp \\ &\quad (.78)\ (.52)\qquad (.045)\qquad (.00004) \\ R^2 &= .100,\ \overline{R}^2 = .098.\end{aligned}$$

$$\widehat{prate} = 97.32 + 5.02\,mrate + .314\,age - 2.66\log(totemp)$$
$$(1.95)\quad (0.51)\qquad (.044)\qquad (.28)$$
$$R^2 = .144,\ \overline{R}^2 = .142.$$

$$\widehat{prate} = 80.62 + 5.34\,mrate + .290\,age - .00043\,totemp$$
$$(.78)\quad (.52)\qquad (.045)\qquad (.00009)$$
$$+ .0000000039\,totemp^2$$
$$(.00000000010)$$
$$R^2 = .108,\ \overline{R}^2 = .106.$$

Which of these three models do you prefer? Why?

8 Suppose we want to estimate the effects of alcohol consumption (*alcohol*) on college grade point average (*colGPA*). In addition to collecting information on grade point averages and alcohol usage, we also obtain attendance information (say, percentage of lectures attended, called *attend*). A standardized test score (say, *SAT*) and high school GPA (*hsGPA*) are also available.

(i) Should we include *attend* along with *alcohol* as explanatory variables in a multiple regression model? (Think about how you would interpret β_{alcohol}.)

(ii) Should *SAT* and *hsGPA* be included as explanatory variables? Explain.

9 If we start with (6.38) under the CLM assumptions, assume large n, and ignore the estimation error in the $\hat{\beta}_j$, a 95% prediction interval for y^0 is $[\exp(-1.96\hat{\sigma})\exp(\widehat{logy^0}),\ \exp(1.96\hat{\sigma})\exp(\widehat{logy^0})]$. The point prediction for y^0 is $\hat{y}^0 = \exp(\hat{\sigma}^2/2)\exp(\widehat{logy^0})$.

(i) For what values of $\hat{\sigma}$ will the point prediction be in the 95% prediction interval? Does this condition seem likely to hold in most applications?

(ii) Verify that the condition from part (i) is satisfied in the CEO salary example.

10 The following two equations were estimated using the data in MEAPSINGLE. The key explanatory variable is *lexppp*, the log of expenditures per student at the school level.

$$\widehat{math4} = 24.49 + 9.01\,lexppp - .422\,free - .752\,lmedinc - .274\,pctsgle$$
$$(59.24)\quad (4.04)\qquad (.071)\qquad (5.358)\qquad (.161)$$
$$n = 229,\ R^2 = .472,\ \overline{R}^2 = .462.$$

$$\widehat{math4} = 149.38 + 1.93\,lexppp - .060\,free - 10.78\,lmedinc - .397\,pctsgle + .667\,read4$$
$$(41.70)\quad (2.82)\qquad (.054)\qquad (3.76)\qquad (.111)\qquad (.042)$$
$$n = 229,\ R^2 = .749,\ \overline{R}^2 = .743.$$

(i) If you are a policy maker trying to estimate the causal effect of per-student spending on math test performance, explain why the first equation is more relevant than the second. What is the estimated effect of a 10% increase in expenditures per student?

(ii) Does adding *read4* to the regression have strange effects on coefficients and statistical significance other than β_{lexppp}?

(iii) How would you explain to someone with only basic knowledge of regression why, in this case, you prefer the equation with the smaller adjusted R-squared?

11 Consider the equation

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$
$$\mathrm{E}(u|x) = 0,$$

where the explanatory variable x has a standard normal distribution in the population. In particular, $\mathrm{E}(x) = 0$, $\mathrm{E}(x^2) = \mathrm{Var}(x) = 1$, and $\mathrm{E}(x^3) = 0$. This last condition holds because the standard normal

distribution is symmetric about zero. We want to study what we can say about the OLS estimator if we omit x^2 and compute the simple regression estimator of the intercept and slope.

(i) Show that we can write

$$y = \alpha_0 + \beta_1 x + v.$$

where $E(v) = 0$. In particular, find v and the new intercept, α_0.

(ii) Show that $E(v|x)$ depends on x unless $\beta_2 = 0$.

(iii) Show that $\text{Cov}(x, v) = 0$.

(iv) If $\hat{\beta}_1$ is the slope coefficient from regression y_i on x_i, is $\hat{\beta}_1$ consistent for β_1? Is it unbiased? Explain.

(v) Argue that being able to estimate β_1 has some value in the following sense: β_1 is the partial effect of x on $E(y|x)$evaluated at $x = 0$, the average value of x.

(vi) Explain why being able to consistently estimate β_1 and β_2 is more valuable than just estimating β_1.

Computer Exercises

C1 Use the data in KIELMC, only for the year 1981, to answer the following questions. The data are for houses that sold during 1981 in North Andover, Massachusetts; 1981 was the year construction began on a local garbage incinerator.

(i) To study the effects of the incinerator location on housing price, consider the simple regression model

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u,$$

where *price* is housing price in dollars and *dist* is distance from the house to the incinerator measured in feet. Interpreting this equation causally, what sign do you expect for β_1 if the presence of the incinerator depresses housing prices? Estimate this equation and interpret the results.

(ii) To the simple regression model in part (i), add the variables log(*intst*), log(*area*), log(*land*), *rooms*, *baths*, and *age*, where *intst* is distance from the home to the interstate, *area* is square footage of the house, *land* is the lot size in square feet, *rooms* is total number of rooms, *baths* is number of bathrooms, and *age* is age of the house in years. Now, what do you conclude about the effects of the incinerator? Explain why (i) and (ii) give conflicting results.

(iii) Add $[\log(intst)]^2$ to the model from part (ii). Now what happens? What do you conclude about the importance of functional form?

(iv) Is the square of log(*dist*) significant when you add it to the model from part (iii)?

C2 Use the data in WAGE1 for this exercise.

(i) Use OLS to estimate the equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

and report the results using the usual format.

(ii) Is $exper^2$ statistically significant at the 1% level?

(iii) Using the approximation

$$\%\Delta\widehat{wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

(iv) At what value of *exper* does additional experience actually lower predicted log(*wage*)? How many people have more experience in this sample?

C3 Consider a model in which the return to education depends upon the amount of work experience (and vice versa):

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper + u.$$

(i) Show that the return to another year of education (in decimal form), holding *exper* fixed, is $\beta_1 + \beta_3 exper$.

(ii) State the null hypothesis that the return to education does not depend on the level of *exper*. What do you think is the appropriate alternative?

(iii) Use the data in WAGE2 to test the null hypothesis in (ii) against your stated alternative.

(iv) Let θ_1 denote the return to education (in decimal form), when $exper = 10$: $\theta_1 = \beta_1 + 10\beta_3$. Obtain $\hat{\theta}_1$ and a 95% confidence interval for θ_1. (*Hint:* Write $\beta_1 = \theta_1 - 10\beta_3$ and plug this into the equation; then rearrange. This gives the regression for obtaining the confidence interval for θ_1.)

C4 Use the data in GPA2 for this exercise.

(i) Estimate the model

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u,$$

where *hsize* is the size of the graduating class (in hundreds), and write the results in the usual form. Is the quadratic term statistically significant?

(ii) Using the estimated equation from part (i), what is the "optimal" high school size? Justify your answer.

(iii) Is this analysis representative of the academic performance of *all* high school seniors? Explain.

(iv) Find the estimated optimal high school size, using log(*sat*) as the dependent variable. Is it much different from what you obtained in part (ii)?

C5 Use the housing price data in HPRICE1 for this exercise.

(i) Estimate the model

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + u$$

and report the results in the usual OLS format.

(ii) Find the predicted value of log(*price*), when $lotsize = 20{,}000$, $sqrft = 2{,}500$, and $bdrms = 4$. Using the methods in Section 6-4, find the predicted value of *price* at the same values of the explanatory variables.

(iii) For explaining variation in *price*, decide whether you prefer the model from part (i) or the model

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u.$$

C6 Use the data in VOTE1 for this exercise.

(i) Consider a model with an interaction between expenditures:

$$voteA = \beta_0 + \beta_1 prtystrA + \beta_2 expendA + \beta_3 expendB + \beta_4 expendA \cdot expendB + u.$$

What is the partial effect of *expendB* on *voteA*, holding *prtystrA* and *expendA* fixed? What is the partial effect of *expendA* on *voteA*? Is the expected sign for β_4 obvious?

(ii) Estimate the equation in part (i) and report the results in the usual form. Is the interaction term statistically significant?

(iii) Find the average of *expendA* in the sample. Fix *expendA* at 300 (for \$300,000). What is the estimated effect of another \$100,000 spent by Candidate B on *voteA*? Is this a large effect?

(iv) Now fix *expendB* at 100. What is the estimated effect of $\Delta expendA = 100$ on *voteA*? Does this make sense?

(v) Now, estimate a model that replaces the interaction with *shareA*, Candidate A's percentage share of total campaign expenditures. Does it make sense to hold both *expendA* and *expendB* fixed, while changing *shareA*?

(vi) (Requires calculus) In the model from part (v), find the partial effect of *expendB* on *voteA*, holding *prtystrA* and *expendA* fixed. Evaluate this at $expendA = 300$ and $expendB = 0$ and comment on the results.

C7 Use the data in ATTEND for this exercise.

(i) In the model of Example 6.3, argue that

$$\Delta stndfnl/\Delta priGPA \approx \beta_2 + 2\beta_4 priGPA + \beta_6 atndrte.$$

Use equation (6.19) to estimate the partial effect when $priGPA = 2.59$ and $atndrte = 82$. Interpret your estimate.

(ii) Show that the equation can be written as

$$stndfnl = \theta_0 + \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2.59)^2 + \beta_5 ACT^2 + \beta_6 priGPA(atndrte - 82) + u,$$

where $\theta_2 = \beta_2 + 2\beta_4(2.59) + \beta_6(82)$. (Note that the intercept has changed, but this is unimportant.) Use this to obtain the standard error of $\hat{\theta}_2$ from part (i).

(iii) Suppose that, in place of $priGPA(atndrte - 82)$, you put $(priGPA - 2.59) \cdot (atndrte - 82)$. Now how do you interpret the coefficients on *atndrte* and *priGPA*?

C8 Use the data in HPRICE1 for this exercise.

(i) Estimate the model

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

and report the results in the usual form, including the standard error of the regression. Obtain predicted price, when we plug in $lotsize = 10{,}000$, $sqrft = 2{,}300$, and $bdrms = 4$; round this price to the nearest dollar.

(ii) Run a regression that allows you to put a 95% confidence interval around the predicted value in part (i). Note that your prediction will differ somewhat due to rounding error.

(iii) Let $price^0$ be the unknown future selling price of the house with the characteristics used in parts (i) and (ii). Find a 95% CI for $price^0$ and comment on the width of this confidence interval.

C9 The data set NBASAL contains salary information and career statistics for 269 players in the National Basketball Association (NBA).

(i) Estimate a model relating points-per-game (*points*) to years in the league (*exper*), *age*, and years played in college (*coll*). Include a quadratic in *exper*; the other variables should appear in level form. Report the results in the usual way.

(ii) Holding college years and age fixed, at what value of experience does the next year of experience actually reduce points-per-game? Does this make sense?

(iii) Why do you think *coll* has a negative and statistically significant coefficient? (*Hint:* NBA players can be drafted before finishing their college careers and even directly out of high school.)

(iv) Add a quadratic in *age* to the equation. Is it needed? What does this appear to imply about the effects of age, once experience and education are controlled for?

(v) Now regress log(*wage*) on *points*, *exper*, $exper^2$, *age*, and *coll*. Report the results in the usual format.

(vi) Test whether *age* and *coll* are jointly significant in the regression from part (v). What does this imply about whether age and education have separate effects on wage, once productivity and seniority are accounted for?

C10 Use the data in BWGHT2 for this exercise.

(i) Estimate the equation

$$\log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + u$$

by OLS, and report the results in the usual way. Is the quadratic term significant?

(ii) Show that, based on the equation from part (i), the number of prenatal visits that maximizes log(*bwght*) is estimated to be about 22. How many women had at least 22 prenatal visits in the sample?

(iii) Does it make sense that birth weight is actually predicted to decline after 22 prenatal visits? Explain.

(iv) Add mother's age to the equation, using a quadratic functional form. Holding *npvis* fixed, at what mother's age is the birth weight of the child maximized? What fraction of women in the sample are older than the "optimal" age?

(v) Would you say that mother's age and number of prenatal visits explain a lot of the variation in log(*bwght*)?

(vi) Using quadratics for both *npvis* and *age*, decide whether using the natural log or the level of *bwght* is better for predicting *bwght*.

C11 Use APPLE to verify some of the claims made in Section 6-3.

(i) Run the regression *ecolbs* on *ecoprc*, *regprc* and report the results in the usual form, including the *R*-squared and adjusted *R*-squared. Interpret the coefficients on the price variables and comment on their signs and magnitudes.

(ii) Are the price variables statistically significant? Report the *p*-values for the individual *t* tests.

(iii) What is the range of fitted values for *ecolbs*? What fraction of the sample reports *ecolbs* = 0? Comment.

(iv) Do you think the price variables together do a good job of explaining variation in *ecolbs*? Explain.

(v) Add the variables *faminc*, *hhsize* (household size), *educ*, and *age* to the regression from part (i). Find the *p*-value for their joint significance. What do you conclude?

(vi) Run separate simple regressions of *ecolbs* on *ecoprc* and then *ecolbs* on *regprc*. How do the simple regression coefficients compare with the multiple regression from part (i)? Find the correlation coefficient between *ecoprc* and *regprc* to help explain your findings.

(vii) Consider a model that adds family income and the quantity demanded for regular apples:

$$ecolbs = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc + \beta_4 reglbs + u.$$

From basic economic theory, which explanatory variable does not belong to the equation? When you drop the variables one at a time, do the sizes of the adjusted *R*-squareds affect your answer?

C12 Use the subset of 401KSUBS with *fsize* = 1; this restricts the analysis to single-person households; see also Computer Exercise C8 in Chapter 4.

(i) The youngest age in the sample is 25. How many people are 25 years old?

(ii) In the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + u,$$

what is the literal interpretation of β_2? By itself, is it of much interest?

(iii) Estimate the model from part (ii) and report the results in standard form. Are you concerned that the coefficient on *age* is negative? Explain.

(iv) Because the youngest people in the sample are 25, it makes sense to think that, for a given level of income, the lowest average amount of net total financial assets is at age 25. Recall that the partial effect of *age* on *nettfa* is $\beta_2 + 2\beta_3 age$, so the partial effect at age 25 is $\beta_2 + 2\beta_3(25) = \beta_2 + 50\beta_3$; call this θ_2. Find $\hat{\theta}_2$ and obtain the two-sided *p*-value for testing $H_0: \theta_2 = 0$. You should conclude that $\hat{\theta}_2$ is small and very statistically insignificant. [*Hint:* One way to do this is to estimate the model $nettfa = \alpha_0 + \beta_1 inc + \theta_2 age + \beta_3(age - 25)^2 + u$, where the intercept, α_0 is different from β_0. There are other ways, too.]

(v) Because the evidence against $H_0: \theta_2 = 0$ is very weak, set it to zero and estimate the model

$$nettfa = \alpha_0 + \beta_1 inc + \beta_3(age - 25)^2 + u.$$

In terms of goodness-of-fit, does this model fit better than that in part (ii)?

(vi) For the estimated equation in part (v), set $inc = 30$ (roughly, the average value) and graph the relationship between *nettfa* and *age*, but only for $age \geq 25$. Describe what you see.

(vii) Check to see whether including a quadratic in *inc* is necessary.

C13 Use the data in MEAP00 to answer this question.

(i) Estimate the model

$$math4 = \beta_0 + \beta_2 lexppp + \beta_2 lenroll + \beta_3 lunch + u$$

by OLS, and report the results in the usual form. Is each explanatory variable statistically significant at the 5% level?

(ii) Obtain the fitted values from the regression in part (i). What is the range of fitted values? How does it compare with the range of the actual data on *math4*?

(iii) Obtain the residuals from the regression in part (i). What is the building code of the school that has the largest (positive) residual? Provide an interpretation of this residual.

(iv) Add quadratics of all explanatory variables to the equation, and test them for joint significance. Would you leave them in the model?

(v) Returning to the model in part (i), divide the dependent variable and each explanatory variable by its sample standard deviation, and rerun the regression. (Include an intercept unless you also first subtract the mean from each variable.) In terms of standard deviation units, which explanatory variable has the largest effect on the math pass rate?

C14 Use the data in BENEFITS to answer this question. It is a school-level data set at the K–5 level on average teacher salary and benefits. See Example 4.10 for background.

(i) Regress *lavgsal* on *bs* and report the results in the usual form. Can you reject $H_0: \beta_{bs} = 0$ against a two-sided alternative? Can you reject $H_0: \beta_{bs} = -1$ against $H_1: \beta_{bs} > -1$? Report the *p*-values for both tests.

(ii) Define $lbs = \log(bs)$. Find the range of values for *lbs* and find its standard deviation. How do these compare to the range and standard deviation for *bs*?

(iii) Regress *lavgsal* on *lbs*. Does this fit better than the regression from part (i)?

(iv) Estimate the equation

$$lavgsal = \beta_0 + \beta_1 bs + \beta_2 lenroll + \beta_3 lstaff + \beta_4 lunch + u$$

and report the results in the usual form. What happens to the coefficient on *bs*? Is it now statistically different from zero?

(v) Interpret the coefficient on *lstaff*. Why do you think it is negative?

(vi) Add $lunch^2$ to the equation from part (iv). Is it statistically significant? Compute the turning point (minimum value) in the quadratic, and show that it is within the range of the observed data on *lunch*. How many values of *lunch* are higher than the calculated turning point?

(vii) Based on the findings from part (vi), describe how teacher salaries relate to school poverty rates. In terms of teacher salary, and holding other factors fixed, is it better to teach at a school with $lunch = 0$ (no poverty), $lunch = 50$, or $lunch = 100$ (all kids eligible for the free lunch program)?

APPENDIX 6A

6A. A Brief Introduction to Bootstrapping

In many cases where formulas for standard errors are hard to obtain mathematically, or where they are thought not to be very good approximations to the true sampling variation of an estimator, we can rely on a **resampling method**. The general idea is to treat the observed data as a population that we can draw samples from. The most common resampling method is the **bootstrap**. (There are actually several versions of the bootstrap, but the most general, and most easily applied, is called the *nonparametric bootstrap*, and that is what we describe here.)

Suppose we have an estimate, $\hat{\theta}$, of a population parameter, θ. We obtained this estimate, which could be a function of OLS estimates (or estimates that we cover in later chapters), from a random sample of size n. We would like to obtain a standard error for $\hat{\theta}$ that can be used for constructing t statistics or confidence intervals. Remarkably, we can obtain a valid standard error by computing the estimate from different random samples drawn from the original data.

Implementation is easy. If we list our observations from 1 through n, we draw n numbers randomly, with replacement, from this list. This produces a new data set (of size n) that consists of the original data, but with many observations appearing multiple times (except in the rather unusual case that we resample the original data). Each time we randomly sample from the original data, we can estimate θ using the same procedure that we used on the original data. Let $\hat{\theta}^{(b)}$ denote the estimate from bootstrap sample b. Now, if we repeat the resampling and estimation m times, we have m new estimates, $\{\hat{\theta}^{(b)}: b = 1, 2, \ldots, m\}$. The **bootstrap standard error** of $\hat{\theta}$ is just the sample standard deviation of the $\hat{\theta}^{(b)}$, namely,

$$\text{bse}(\hat{\theta}) = \left[(m-1)^{-1}\sum_{b=1}^{m}(\hat{\theta}^{(b)} - \bar{\hat{\theta}})^2\right]^{1/2}, \qquad [6.50]$$

where $\bar{\hat{\theta}}$ is the average of the bootstrap estimates.

If obtaining an estimate of θ on a sample of size n requires little computational time, as in the case of OLS and all the other estimators we encounter in this text, we can afford to choose m—the number of bootstrap replications—to be large. A typical value is $m = 1{,}000$, but even $m = 500$ or a somewhat smaller value can produce a reliable standard error. Note that the size of m—the number of times we resample the original data—has nothing to do with the sample size, n. (For certain estimation problems beyond the scope of this text, a large n can force one to do fewer bootstrap replications.) Many statistics and econometrics packages have built-in bootstrap commands, and this makes the calculation of bootstrap standard errors simple, especially compared with the work often required to obtain an analytical formula for an asymptotic standard error.

One can actually do better in most cases by using the bootstrap sample to compute p-values for t statistics (and F statistics), or for obtaining confidence intervals, rather than obtaining a bootstrap standard error to be used in the construction of t statistics or confidence intervals. See Horowitz (2001) for a comprehensive treatment.

CHAPTER 7

Multiple Regression Analysis with Qualitative Information

Almost all of our discussion in the previous chapters has focused on the case where the dependent and independent variables in our multiple regression models have quantitative meaning. Just a few examples include hourly wage rate, years of education, college grade point average, amount of air pollution, level of firm sales, and number of arrests. In each case, the magnitude of the variable conveys useful information. In some cases, we take the natural log and then the coefficients can be turned into percentage changes.

In Section 2-7 we introduced the notion of a binary (or dummy) explanatory variable, and we discussed how simple regression on a binary variable can be used to evaluate randomized interventions. We showed how to extend program evaluation to the multiple regression case in Sections 3-7e and 4-7 when it is necessary to account for observed differences between the control and treatment groups.

The purpose of this chapter is to provide a comprehensive analysis of how to include qualitative factors into regression models. In addition to indicators of participating in a program, or being subjected to a new policy, the race or ethnicity of an individual, marital status, the industry of a firm (manufacturing, retail, and so on), and the region in the United States where a city is located (South, North, West, and so on) are common examples of qualitative factors.

After we discuss the appropriate ways to describe qualitative informationinSection7-1, we show how qualitative explanatory variables can be easily incorporated into multiple regression models in Sections 7-2, 7-3, and 7-4. These sections cover almost all of the popular ways that qualitative

independent variables are used in cross-sectional regression analysis, including creating interactions among qualitative variables and between qualitative and quantitative variables.

In Section 7-5, we discuss the case where our dependent variable is binary, which is a particular kind of qualitative dependent variable. The multiple regression model is called the linear probability model (LPM), and the coefficients can be interpreted as changes in a probability. While much maligned by some econometricians, the simplicity of the LPM makes it useful in many empirical contexts. We will describe drawbacks of the LPM in Section 7-5, but they are often secondary in empirical work.

Section 7.6 reconsiders policy analysis, including the potential outcomes perspective, and proposes a flexible regression approach for estimating the effects of interventions. Section 7.7 is a short section that explains how to interpret multiple regression estimates when y is a discrete variable that has quantitative meeting.

This chapter does not assume you have read the material on potential outcomes and policy analysis in Chapters 2, 3, and 4, and so it stands alone as a dicussion of how to incorporate qualitative information into regression.

7-1 Describing Qualitative Information

Qualitative factors often come in the form of binary information: a person is female or male; a person does or does not own a personal computer; a firm offers a certain kind of employee pension plan or it does not; a state administers capital punishment or it does not. In all of these examples, the relevant information can be captured by defining a **binary variable** or a **zero-one variable**. In econometrics, binary variables are most commonly called **dummy variables**, although this name is not especially descriptive.

GOING FURTHER 7.1

Suppose that, in a study comparing election outcomes between Democratic and Republican candidates, you wish to indicate the party of each candidate. Is a name such as *party* a wise choice for a binary variable in this case? What would be a better name?

In defining a dummy variable, we must decide which event is assigned the value one and which is assigned the value zero. For example, in a study of individual wage determination, we might define *female* to be a binary variable taking on the value one for females and the value zero for males. The name in this case indicates the event with the value one. The same information is captured by defining *male* to be one if the person is male and zero if the person is female. Either of these is better than using *gender* because this name does not make it clear when the dummy variable is one: does $gender = 1$ correspond to male or female? What we call our variables is unimportant for getting regression results, but it always helps to choose names that clarify equations and expositions.

Suppose in the wage example that we have chosen the name *female* to indicate gender. Further, we define a binary variable *married* to equal one if a person is married and zero if otherwise. Table 7.1 gives a partial listing of a wage data set that might result. We see that Person 1 is female and not married, Person 2 is female and married, Person 3 is male and not married, and so on.

Why do we use the values zero and one to describe qualitative information? In a sense, these values are arbitrary: any two different values would do. The real benefit of capturing qualitative information using zero-one variables is that it leads to regression models where the parameters have very natural interpretations, as we will see now.

TABLE 7.1 A Partial Listing of the Data in WAGE1

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

7-2 A Single Dummy Independent Variable

How do we incorporate binary information into regression models? In the simplest case, with only a single dummy explanatory variable, we just add it as an independent variable in the equation. For example, consider the following simple model of hourly wage determination:

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u. \qquad [7.1]$$

We use δ_0 as the parameter on *female* in order to highlight the interpretation of the parameters multiplying dummy variables; later, we will use whatever notation is most convenient.

In model (7.1), only two observed factors affect wage: gender and education. Because $female = 1$ when the person is female, and $female = 0$ when the person is male, the parameter δ_0 has the following interpretation: δ_0 is the difference in hourly wage between females and males, *given* the same amount of education (and the same error term u). Thus, the coefficient δ_0 determines whether there is discrimination against women: if $\delta_0 < 0$, then for the same level of other factors, women earn less than men on average.

In terms of expectations, if we assume the zero conditional mean assumption $\mathrm{E}(u|female, educ) = 0$, then

$$\delta_0 = \mathrm{E}(wage|female = 1, educ) - \mathrm{E}(wage|female = 0, educ).$$

Because $female = 1$ corresponds to females and $female = 0$ corresponds to males, we can write this more simply as

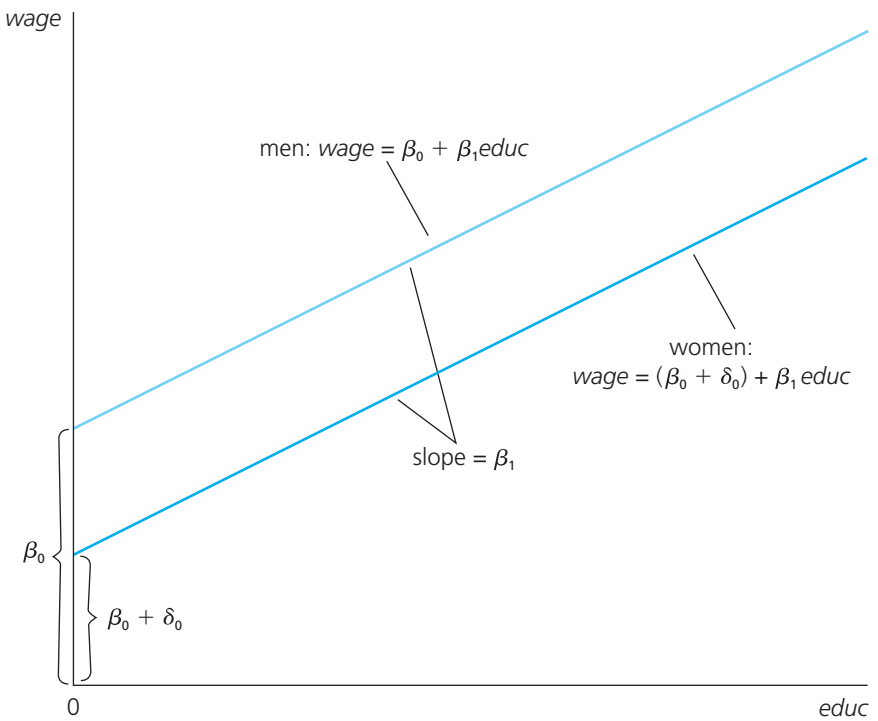
$$\delta_0 = \mathrm{E}(wage|female, educ) - \mathrm{E}(wage|male, educ). \qquad [7.2]$$

The key here is that the level of education is the same in both expectations; the difference, δ_0, is due to gender only.

The situation can be depicted graphically as an **intercept shift** between males and females. In Figure 7.1, the case $\delta_0 < 0$ is shown, so that men earn a fixed amount more per hour than women. The difference does not depend on the amount of education, and this explains why the wage-education profiles for women and men are parallel.

At this point, you may wonder why we do not also include in (7.1) a dummy variable, say *male*, which is one for males and zero for females. This would be redundant. In (7.1), the intercept for males is β_0, and the intercept for females is $\beta_0 + \delta_0$. Because there are just two groups, we only need two different intercepts. This means that, in addition to β_0, we need to use only *one* dummy variable; we have chosen to include the dummy variable for females. Using two dummy variables would introduce

FIGURE 7.1 Graph of $wage = \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$.



perfect collinearity because $female + male = 1$, which means that *male* is a perfect linear function of *female*. Including dummy variables for both genders is the simplest example of the so-called **dummy variable trap**, which arises when too many dummy variables describe a given number of groups. We will discuss this problem in detail later.

In (7.1), we have chosen males to be the **base group** or **benchmark group**, that is, the group against which comparisons are made. This is why β_0 is the intercept for males, and δ_0 is the *difference* in intercepts between females and males. We could choose females as the base group by writing the model as

$$wage = \alpha_0 + \gamma_0 male + \beta_1 educ + u,$$

where the intercept for females is α_0 and the intercept for males is $\alpha_0 + \gamma_0$; this implies that $\alpha_0 = \beta_0 + \delta_0$ and $\alpha_0 + \gamma_0 = \beta_0$. In any application, it does not matter how we choose the base group, but it is important to keep track of which group is the base group.

Some researchers prefer to drop the overall intercept in the model and to include dummy variables for each group. The equation would then be $wage = \beta_0 male + \alpha_0 female + \beta_1 educ + u$, where the intercept for men is β_0 and the intercept for women is α_0. There is no dummy variable trap in this case because we do not have an overall intercept. However, this formulation has little to offer, because testing for a difference in the intercepts is more difficult, and there is no generally agreed upon way to compute R-squared in regressions without an intercept. Therefore, we will always include an overall intercept for the base group.

Nothing much changes when more explanatory variables are involved. Taking males as the base group, a model that controls for experience and tenure in addition to education is

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u. \qquad [7.3]$$

If *educ*, *exper*, and *tenure* are all relevant productivity characteristics, the null hypothesis of *no* difference between men and women is H_0: $\delta_0 = 0$. The alternative that there is discrimination against women is H_1: $\delta_0 < 0$.

How can we actually test for wage discrimination? The answer is simple: just estimate the model by OLS, *exactly* as before, and use the usual *t* statistic. Nothing changes about the mechanics of OLS or the statistical theory when some of the independent variables are defined as dummy variables. The only difference with what we have done up until now is in the interpretation of the coefficient on the dummy variable.

EXAMPLE 7.1 Hourly Wage Equation

Using the data in WAGE1, we estimate model (7.3). For now, we use *wage*, rather than log(*wage*), as the dependent variable:

$$\widehat{wage} = -1.57 - 1.81\,female + .572\,educ + 0.25\,exper + .141\,tenure \quad [7.4]$$
$$(.72)\quad (.26)\qquad (.049)\qquad (.012)\qquad (.021)$$
$$n = 526,\ R^2 = .364.$$

The negative intercept—the intercept for men, in this case—is not very meaningful because no one has zero values for all of *educ*, *exper*, and *tenure* in the sample. The coefficient on *female* is interesting because it measures the average difference in hourly wage between a man and a woman who have the *same* levels of *educ*, *exper*, and *tenure*. If we take a woman and a man with the same levels of education, experience, and tenure, the woman earns, on average, $1.81 less per hour than the man. (Recall that these are 1976 wages.)

It is important to remember that, because we have performed multiple regression and controlled for *educ*, *exper*, and *tenure*, the $1.81 wage differential cannot be explained by different average levels of education, experience, or tenure between men and women. We can conclude that the differential of $1.81 is due to gender or factors associated with gender that we have not controlled for in the regression. [In 2013 dollars, the wage differential is about $4.09(1.81) \approx 7.40$.]

It is informative to compare the coefficient on *female* in equation (7.4) to the estimate we get when all other explanatory variables are dropped from the equation:

$$\widehat{wage} = 7.10 - 2.51\,female \quad [7.5]$$
$$(.21)\quad (.30)$$
$$n = 526,\ R^2 = .116.$$

As discussed in Section 2.7, the coefficients in (7.5) have a simple interpretation. The intercept is the average wage for men in the sample (let *female* = 0), so men earn $7.10 per hour on average. The coefficient on *female* is the difference in the average wage between women and men. Thus, the average wage for women in the sample is $7.10 - 2.51 = 4.59$, or $4.59 per hour. (Incidentally, there are 274 men and 252 women in the sample.)

Equation (7.5) provides a simple way to carry out a *comparison-of-means test* between the two groups, which in this case are men and women. The estimated difference, −2.51, has a *t* statistic of −8.37, which is very statistically significant (and, of course, $2.51 is economically large as well). Generally, simple regression on a constant and a dummy variable is a straightforward way to compare the means of two groups. For the usual *t* test to be valid, we must assume that the homoskedasticity assumption holds, which means that the population variance in wages for men is the same as that for women.

The estimated wage differential between men and women is larger in (7.5) than in (7.4) because (7.5) does not control for differences in education, experience, and tenure, and these are lower, on average, for women than for men in this sample. Equation (7.4) gives a more reliable estimate of the ceteris paribus gender wage gap; it still indicates a very large differential.

In many cases, dummy independent variables reflect choices of individuals or other economic units (as opposed to something predetermined, such as gender). In such situations, the matter of causality is again a central issue. In the following example, we would like to know whether personal computer ownership *causes* a higher college grade point average.

EXAMPLE 7.2 **Effects of Computer Ownership on College GPA**

In order to determine the effects of computer ownership on college grade point average, we estimate the model

$$colGPA = \beta_0 + \delta_0 PC + \beta_1 hsGPA + \beta_2 ACT + u,$$

where the dummy variable *PC* equals one if a student owns a personal computer and zero otherwise. There are various reasons PC ownership might have an effect on *colGPA*. A student's schoolwork might be of higher quality if it is done on a computer, and time can be saved by not having to wait at a computer lab. Of course, a student might be more inclined to play computer games or surf the Internet if he or she owns a PC, so it is not obvious that δ_0 is positive. The variables *hsGPA* (high school GPA) and *ACT* (achievement test score) are used as controls: it could be that stronger students, as measured by high school *GPA* and *ACT* scores, are more likely to own computers. We control for these factors because we would like to know the average effect on *colGPA* if a student is picked at random and given a personal computer.

Using the data in GPA1, we obtain

$$\widehat{colGPA} = 1.26 + .157\, PC + .447\, hsGPA + .0087\, ACT \quad [7.6]$$
$$(.33)\ (.057)\quad (.094)\quad (.0105)$$
$$n = 141,\ R^2 = .219.$$

This equation implies that a student who owns a PC has a predicted GPA about .16 points higher than a comparable student without a PC (remember, both *colGPA* and *hsGPA* are on a four-point scale). The effect is also very statistically significant, with $t_{PC} = .157/.057 \approx 2.75$.

What happens if we drop *hsGPA* and *ACT* from the equation? Clearly, dropping the latter variable should have very little effect, as its coefficient and *t* statistic are very small. But *hsGPA* is very significant, and so dropping it could affect the estimate of β_{PC}. Regressing *colGPA* on *PC* gives an estimate on *PC* equal to about .170, with a standard error of .063; in this case, $\hat{\beta}_{PC}$ and its *t* statistic do not change by much.

In the exercises at the end of the chapter, you will be asked to control for other factors in the equation to see if the computer ownership effect disappears, or if it at least gets notably smaller.

Each of the previous examples can be viewed as having relevance for **policy analysis**. In the first example, we were interested in gender discrimination in the workforce. In the second example, we were concerned with the effect of computer ownership on college performance. A special case of policy analysis is **program evaluation**, where we would like to know the effect of economic or social programs on individuals, firms, neighborhoods, cities, and so on.

In the simplest case, there are two groups of subjects. The **control group** does not participate in the program. The **experimental group** or **treatment group** does take part in the program. These names come from literature in the experimental sciences, and they should not be taken literally. Except in rare cases, the choice of the control and treatment groups is not random. However, in some cases, multiple regression analysis can be used to control for enough other factors in order to estimate the causal effect of the program.

EXAMPLE 7.3 **Effects of Training Grants on Hours of Training**

Using the 1988 data for Michigan manufacturing firms in JTRAIN, we obtain the following estimated equation:

$$\widehat{hrsemp} = 46.67 + 26.25\,grant - .98\log(sales) - 6.07\log(employ) \quad [7.7]$$
$$(43.41)\quad (5.59)\qquad (3.54)\qquad\quad (3.88)$$
$$n = 105,\ R^2 = .237.$$

The dependent variable is hours of training per employee, at the firm level. The variable *grant* is a dummy variable equal to one if the firm received a job training grant for 1988, and zero otherwise. The variables *sales* and *employ* represent annual sales and number of employees, respectively. We cannot enter *hrsemp* in logarithmic form because *hrsemp* is zero for 29 of the 105 firms used in the regression.

The variable *grant* is very statistically significant, with $t_{grant} = 4.70$. Controlling for sales and employment, firms that received a grant trained each worker, on average, 26.25 hours more. Because the average number of hours of per worker training in the sample is about 17, with a maximum value of 164, *grant* has a large effect on training, as is expected.

The coefficient on log(*sales*) is small and very insignificant. The coefficient on log(*employ*) means that, if a firm is 10% larger, it trains its workers about .61 hour less. Its t statistic is -1.56, which is only marginally statistically significant.

As with any other independent variable, we should ask whether the measured effect of a qualitative variable is causal. In equation (7.7), is the difference in training between firms that receive grants and those that do not due to the grant, or is grant receipt simply an indicator of something else? It might be that the firms receiving grants would have, on average, trained their workers more even in the absence of a grant. Nothing in this analysis tells us whether we have estimated a causal effect; we must know how the firms receiving grants were determined. We can only hope we have controlled for as many factors as possible that might be related to whether a firm received a grant and to its levels of training.

In Section 7.6 we return to policy analysis using binary indicators, including obtaining a more flexible framework in the context of potential outcomes. These themes reappear in the remainder of the text.

7-2a Interpreting Coefficients on Dummy Explanatory Variables When the Dependent Variable Is log(*y*)

A common specification in applied work has the dependent variable appearing in logarithmic form, with one or more dummy variables appearing as independent variables. How do we interpret the dummy variable coefficients in this case? Not surprisingly, the coefficients have a *percentage* interpretation.

EXAMPLE 7.4 **Housing Price Regression**

Using the data in HPRICE1, we obtain the equation

$$\widehat{\log(price)} = -1.35 + .168\log(lotsize) + .707\log(sqrft)$$
$$(.65)\quad (.038)\qquad\quad (.093)$$
$$+ .027\,bdrms + .054\,colonial \quad [7.8]$$
$$(.029)\qquad (.045)$$
$$n = 88,\ R^2 = .649.$$

All the variables are self-explanatory except colonial, which is a binary variable equal to one if the house is of the colonial style. What does the coefficient on *colonial* mean? For given levels of *lotsize*, *sqrft*, and *bdrms*, the difference in $\widehat{\log(price)}$ between a house of colonial style and that of another style is .054. This means that a colonial-style house is predicted to sell for about 5.4% more, holding other factors fixed.

This example shows that, when $\log(y)$ is the dependent variable in a model, the coefficient on a dummy variable, when multiplied by 100, is interpreted as the percentage difference in y, holding all other factors fixed. When the coefficient on a dummy variable suggests a large proportionate change in y, the exact percentage difference can be obtained exactly as with the semi-elasticity calculation in Section 6-2.

EXAMPLE 7.5 Log Hourly Wage Equation

Let us reestimate the wage equation from Example 7.1, using $\log(wage)$ as the dependent variable and adding quadratics in *exper* and *tenure*:

$$\begin{aligned}\widehat{\log(wage)} &= .417 - .297\,female + .080\,educ + .029\,exper \\ &\qquad (.099)\ (.036)\qquad (.007)\qquad (.005) \\ &\quad - .00058\,exper^2 + .032\,tenure - .00059\,tenure^2 \qquad [7.9] \\ &\qquad (.00010)\qquad (.007)\qquad (.00023) \\ n &= 526,\ R^2 = .441.\end{aligned}$$

Using the same approximation as in Example 7.4, the coefficient on female implies that, for the same levels of *educ*, *exper*, and *tenure*, women earn about 100(.297) = 29.7% less than men. We can do better than this by computing the exact percentage difference in predicted wages. What we want is the proportionate difference in wages between females and males, holding other factors fixed: $(\widehat{wage_F} - \widehat{wage_M})/\widehat{wage_M}$. What we have from (7.9) is

$$\widehat{\log(wage_F)} - \widehat{\log(wage_M)} = -.297.$$

Exponentiating and subtracting one gives

$$(\widehat{wage_F} - \widehat{wage_M})/\widehat{wage_M} = \exp(-.297) - 1 \approx -.257.$$

This more accurate estimate implies that a woman's wage is, on average, 25.7% below a comparable man's wage.

If we had made the same correction in Example 7.4, we would have obtained $\exp(.054) - 1 \approx .0555$, or about 5.6%. The correction has a smaller effect in Example 7.4 than in the wage example because the magnitude of the coefficient on the dummy variable is much smaller in (7.8) than in (7.9).

Generally, if $\hat{\beta}_1$ is the coefficient on a dummy variable, say x_1, when $\log(y)$ is the dependent variable, the exact percentage difference in the predicted y when $x_1 = 1$ versus when $x_1 = 0$ is

$$100 \cdot [\exp(\hat{\beta}_1) - 1]. \qquad [7.10]$$

The estimate $\hat{\beta}_1$ can be positive or negative, and it is important to preserve its sign in computing (7.10).

The logarithmic approximation has the advantage of providing an estimate between the magnitudes obtained by using each group as the base group. In particular, although equation (7.10) gives us a better estimate than $100 \cdot \hat{\beta}_1$ of the percentage by which y for $x_1 = 1$ is greater than y for $x_1 = 0$, (7.10) is not a good estimate if we switch the base group. In Example 7.5, we can estimate

the percentage by which a man's wage exceeds a comparable woman's wage, and this estimate is $100 \cdot [\exp(-\hat{\beta}_1) - 1] = 100 \cdot [\exp(.297) - 1] \approx 34.6$. The approximation, based on $100 \cdot \hat{\beta}_1$, 29.7, is between 25.7 and 34.6 (and close to the middle). Therefore, it makes sense to report that "the difference in predicted wages between men and women is about 29.7%," without having to take a stand on which is the base group.

7-3 Using Dummy Variables for Multiple Categories

We can use several dummy independent variables in the same equation. For example, we could add the dummy variable *married* to equation (7.9). The coefficient on *married* gives the (approximate) proportional differential in wages between those who are and are not married, holding *female*, *educ*, *exper*, and *tenure* fixed. When we estimate this model, the coefficient on *married* (with standard error in parentheses) is .053 (.041), and the coefficient on *female* becomes −.290 (.036). Thus, the "marriage premium" is estimated to be about 5.3%, but it is not statistically different from zero ($t = 1.29$). An important limitation of this model is that the marriage premium is assumed to be the same for men and women; this is relaxed in the following example.

EXAMPLE 7.6 Log Hourly Wage Equation

Let us estimate a model that allows for wage differences among four groups: married men, married women, single men, and single women. To do this, we must select a base group; we choose single men. Then, we must define dummy variables for each of the remaining groups. Call these *marrmale*, *marrfem*, and *singfem*. Putting these three variables into (7.9) (and, of course, dropping *female*, as it is now redundant) gives

$$\begin{aligned}\widehat{\log(wage)} &= .321 + .213\,marrmale - .198\,marrfem \\ &\qquad (.100)\quad (.055)\qquad\quad (.058) \\ &\quad - .110\,singfem + .079\,educ + .027\,exper - .00054\,exper^2 \\ &\qquad (.056)\qquad\quad (.007)\qquad (.005)\qquad\quad (.00011) \\ &\quad + .029\,tenure - .00053\,tenure^2 \\ &\qquad (.007)\qquad\quad (.00023) \\ n &= 526,\ R^2 = .461.\end{aligned} \tag{7.11}$$

All of the coefficients, with the exception of *singfem*, have t statistics well above two in absolute value. The t statistic for *singfem* is about −1.96, which is just significant at the 5% level against a two-sided alternative.

To interpret the coefficients on the dummy variables, we must remember that the base group is single males. Thus, the estimates on the three dummy variables measure the proportionate difference in wage *relative* to single males. For example, married men are estimated to earn about 21.3% more than single men, holding levels of education, experience, and tenure fixed. [The more precise estimate from (7.10) is about 23.7%.] A married woman, on the other hand, earns a predicted 19.8% less than a single man with the same levels of the other variables.

Because the base group is represented by the intercept in (7.11), we have included dummy variables for only three of the four groups. If we were to add a dummy variable for single males to (7.11), we would fall into the dummy variable trap by introducing perfect collinearity. Some regression packages will automatically correct this mistake for you, while others will just tell you there is perfect collinearity. It is best to carefully specify the dummy variables because then we are forced to properly interpret the final model.

Even though single men is the base group in (7.11), we can use this equation to obtain the estimated difference between any two groups. Because the overall intercept is common to all groups, we can ignore that in finding differences. Thus, the estimated proportionate difference between single and married women is $-.110 - (-.198) = .088$, which means that single women earn about 8.8% more than married women. Unfortunately, we cannot use equation (7.11) for testing whether the estimated difference between single and married women is statistically significant. Knowing the standard errors on *marrfem* and *singfem* is not enough to carry out the test (see Section 4-4). The easiest thing to do is to choose one of these groups to be the base group and to reestimate the equation. Nothing substantive changes, but we get the needed estimate and its standard error directly. When we use married women as the base group, we obtain

$$\widehat{\log(wage)} = .123 + .411\,marrmale + .198\,singmale + .088\,singfem + \cdots,$$
$$\qquad\qquad (.106)\quad (.056)\qquad\qquad (.058)\qquad\qquad (.052)$$

where, of course, none of the unreported coefficients or standard errors have changed. The estimate on *singfem* is, as expected, .088. Now, we have a standard error to go along with this estimate. The t statistic for the null that there is no difference in the population between married and single women is $t_{singfem} = .088/.052 \approx 1.69$. This is marginal evidence against the null hypothesis. We also see that the estimated difference between married men and married women is very statistically significant $(t_{marrmale} = 7.34)$.

The previous example illustrates a general principle for including dummy variables to indicate different groups: if the regression model is to have different intercepts for, say, g groups or categories, we need to include $g - 1$ dummy variables in the model along with an intercept. The intercept for the base group is the overall intercept in the model, and the dummy variable coefficient for a particular group represents the estimated difference in intercepts between that group and the base group. Including g dummy variables along with an intercept will result in the dummy variable trap. An alternative is to include g dummy variables and to exclude an overall intercept. Including g dummies without an overall intercept is sometimes useful, but it has two practical drawbacks. First, it makes it more cumbersome to test for differences relative to a base group. Second, regression packages usually change the way R-squared is computed when an overall intercept is not included. In particular, in the formula $R^2 = 1 - \text{SSR/SST}$, the total sum of squares, SST, is replaced with a total sum of squares that does not center y_i about its mean, say, $\text{SST}_0 = \sum_{i=1}^{n} y_i^2$. The resulting R-squared, say $R_0^2 = 1 - \text{SSR/SST}_0$, is sometimes called the **uncentered *R*-squared**. Unfortunately, R_0^2 is rarely suitable as a goodness-of-fit measure. It is always true that $\text{SST}_0 \geq \text{SST}$ with equality only if $\bar{y} = 0$. Often, SST_0 is much larger than SST, which means that R_0^2 is much larger than R^2. For example, if in the previous example we regress log(*wage*) on *marrmale*, *singmale*, *marrfem*, *singfem*, and the other explanatory variables—without an intercept—the reported R-squared from Stata, which is R_0^2, is .948. This high R-squared is an artifact of not centering the total sum of squares in the calculation. The correct R-squared is given in equation (7.11) as .461. Some regression packages, including Stata, have an option to force calculation of the centered R-squared even though an overall intercept has not been included, and using this option is generally a good idea. In the vast majority of cases, any R-squared based on comparing an SSR and SST should have SST computed by centering the y_i about $\bar{y}$. We can think of this SST as the sum of squared residuals obtained if we just use the sample average, $\bar{y}$, to predict each y_i. Surely we are

GOING FURTHER 7.2

In the baseball salary data found in MLB1, players are given one of six positions: *frstbase, scndbase, thrdbase, shrtstop, outfield,* or *catcher*. To allow for salary differentials across position, with outfielders as the base group, which dummy variables would you include as independent variables?

setting the bar pretty low for any model if all we measure is its fit relative to using a constant predictor. For a model without an intercept that fits poorly, it is possible that SSR $>$ SST, which means R^2 would be negative. The uncentered R-squared will always be between zero and one, which likely explains why it is usually the default when an intercept is not estimated in regression models.

7-3a Incorporating Ordinal Information by Using Dummy Variables

Suppose that we would like to estimate the effect of city credit ratings on the municipal bond interest rate (*MBR*). Several financial companies, such as Moody's Investors Service and Standard and Poor's, rate the quality of debt for local governments, where the ratings depend on things like probability of default. (Local governments prefer lower interest rates in order to reduce their costs of borrowing.) For simplicity, suppose that rankings take on the integer values $\{0, 1, 2, 3, 4\}$, with zero being the worst credit rating and four being the best. This is an example of an **ordinal variable**. Call this variable *CR* for concreteness. The question we need to address is: How do we incorporate the variable *CR* into a model to explain *MBR*?

One possibility is to just include *CR* as we would include any other explanatory variable:

$$MBR = \beta_0 + \beta_1 CR + other\ factors,$$

where we do not explicitly show what other factors are in the model. Then β_1 is the percentage point change in *MBR* when *CR* increases by one unit, holding other factors fixed. Unfortunately, it is rather hard to interpret a one-unit increase in *CR*. We know the quantitative meaning of another year of education, or another dollar spent per student, but things like credit ratings typically have only ordinal meaning. We know that a *CR* of four is better than a *CR* of three, but is the difference between four and three the same as the difference between one and zero? If not, then it might not make sense to assume that a one-unit increase in *CR* has a constant effect on *MBR*.

A better approach, which we can implement because *CR* takes on relatively few values, is to define dummy variables for each value of *CR*. Thus, let $CR_1 = 1$ if $CR = 1$, and $CR_1 = 0$ otherwise; $CR_2 = 1$ if $CR = 2$, and $CR_2 = 0$ otherwise; and so on. Effectively, we take the single credit rating and turn it into five categories. Then, we can estimate the model

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + other\ factors. \qquad [7.12]$$

Following our rule for including dummy variables in a model, we include four dummy variables because we have five categories. The omitted category here is a credit rating of zero, and so it is the base group. (This is why we do not need to define a dummy variable for this category.) The coefficients are easy to interpret: δ_1 is the difference in *MBR* (other factors fixed) between a municipality with a credit rating of one and a municipality with a credit rating of zero; δ_2 is the difference in *MBR* between a municipality with a credit rating of two and a municipality with a credit rating of zero; and so on. The movement between each credit rating is allowed to have a different effect, so using (7.12) is much more flexible than simply putting *CR* in as a single variable. Once the dummy variables are defined, estimating (7.12) is straightforward.

GOING FURTHER 7.3

In model (7.12), how would you test the null hypothesis that credit rating has no effect on *MBR*?

Equation (7.12) contains the model with a constant partial effect as a special case. One way to write the three restrictions that imply a constant partial effect is $\delta_2 = 2\delta_1$, $\delta_3 = 3\delta_1$, and $\delta_4 = 4\delta_1$. When we plug these into equation (7.12) and rearrange, we get $MBR = \beta_0 + \delta_1(CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + other\ factors$. Now, the term multiplying δ_1 is simply the original credit rating variable, *CR*. To obtain the *F* statistic for testing the constant partial effect restrictions,

we obtain the unrestricted R-squared from (7.12) and the restricted R-squared from the regression of *MBR* on *CR* and the other factors we have controlled for. The F statistic is obtained as in equation (4.41) with $q = 3$.

EXAMPLE 7.7 **Effects of Physical Attractiveness on Wage**

Hamermesh and Biddle (1994) used measures of physical attractiveness in a wage equation. (The file BEAUTY contains fewer variables but more observations than used by Hamermesh and Biddle. See Computer Exercise C12.) Each person in the sample was ranked by an interviewer for physical attractiveness, using five categories (homely, quite plain, average, good looking, and strikingly beautiful or handsome). Because there are so few people at the two extremes, the authors put people into one of three groups for the regression analysis: average, below average, and above average, where the base group is *average*. Using data from the 1977 Quality of Employment Survey, after controlling for the usual productivity characteristics, Hamermesh and Biddle estimated an equation for men:

$$\widehat{\log(wage)} = \hat{\beta}_0 - \underset{(.046)}{.164\, belavg} + \underset{(.033)}{.016\, abvavg} + other\ factors$$
$$n = 700,\ \bar{R}^2 = .403$$

and an equation for women:

$$\widehat{\log(wage)} = \hat{\beta}_0 - \underset{(.066)}{.124\, belavg} + \underset{(.049)}{.035\, abvavg} + other\ factors$$
$$n = 409,\ \bar{R}^2 = .330.$$

The other factors controlled for in the regressions include education, experience, tenure, marital status, and race; see Table 3 in Hamermesh and Biddle's paper for a more complete list. In order to save space, the coefficients on the other variables are not reported in the paper and neither is the intercept.

For men, those with below average looks are estimated to earn about 16.4% less than an average-looking man who is the same in other respects (including education, experience, tenure, marital status, and race). The effect is statistically different from zero, with $t = -3.57$. Men with above average looks are estimated to earn only 1.6% more than men with average looks, and the effect is not statistically significant $(t < .5)$.

A woman with below average looks earns about 12.4% less than an otherwise comparable average-looking woman, with $t = -1.88$. As was the case for men, the estimate on *abvavg* is much smaller in magnitude and not statistically different from zero.

In related work, Biddle and Hamermesh (1998) revisit the effects of looks on earnings using a more homogeneous group: graduates of a particular law school. The authors continue to find that physical appearance has an effect on annual earnings, something that is perhaps not too surprising among people practicing law.

In some cases, the ordinal variable takes on too many values so that a dummy variable cannot be included for each value. For example, the file LAWSCH85 contains data on median starting salaries for law school graduates. One of the key explanatory variables is the rank of the law school. Because each law school has a different rank, we clearly cannot include a dummy variable for each rank. If we do not wish to put the rank directly in the equation, we can break it down into categories. The following example shows how this is done.

EXAMPLE 7.8 **Effects of Law School Rankings on Starting Salaries**

Define the dummy variables *top10, r11_25, r26_40, r41_60, r61_100* to take on the value unity when the variable *rank* falls into the appropriate range. We let schools ranked below 100 be the base group. The estimated equation is

$$\begin{aligned}\widehat{\log(salary)} &= 9.17 + .700\,top10 + .594\,r11\_25 + .375\,r26\_40\\ &\qquad (.41)\ \ (.053)\qquad\ (.039)\qquad\ (.034)\\ &\quad + .263\,r41\_60 + .132\,r61\_100 + .0057\,LSAT\\ &\qquad (.028)\qquad\ (.021)\qquad\ (.0031)\\ &\quad + .041\,GPA + .036\log(libvol) + .0008\log(cost)\\ &\qquad (.074)\qquad\ (.026)\qquad\ (.0251)\\ n &= 136,\ R^2 = .911,\ \overline{R}^2 = .905.\end{aligned} \tag{7.13}$$

We see immediately that all of the dummy variables defining the different ranks are very statistically significant. The estimate on *r61_100* means that, holding *LSAT*, *GPA*, *libvol*, and *cost* fixed, the median salary at a law school ranked between 61 and 100 is about 13.2% higher than that at a law school ranked below 100. The difference between a top 10 school and a below 100 school is quite large. Using the exact calculation given in equation (7.10) gives $\exp(.700) - 1 \approx 1.014$, and so the predicted median salary is more than 100% higher at a top 10 school than it is at a below 100 school.

As an indication of whether breaking the rank into different groups is an improvement, we can compare the adjusted *R*-squared in (7.13) with the adjusted *R*-squared from including *rank* as a single variable: the former is .905 and the latter is .836, so the additional flexibility of (7.13) is warranted.

Interestingly, once the rank is put into the (admittedly somewhat arbitrary) given categories, all of the other variables become insignificant. In fact, a test for joint significance of *LSAT*, *GPA*, log(*libvol*), and log(*cost*) gives a *p*-value of .055, which is borderline significant. When *rank* is included in its original form, the *p*-value for joint significance is zero to four decimal places.

One final comment about this example: In deriving the properties of ordinary least squares, we assumed that we had a random sample. The current application violates that assumption because of the way *rank* is defined: a school's rank necessarily depends on the rank of the other schools in the sample, and so the data cannot represent independent draws from the population of all law schools. This does not cause any serious problems provided the error term is uncorrelated with the explanatory variables.

7-4 Interactions Involving Dummy Variables

7-4a Interactions among Dummy Variables

Just as variables with quantitative meaning can be interacted in regression models, so can dummy variables. We have effectively seen an example of this in Example 7.6, where we defined four categories based on marital status and gender. In fact, we can recast that model by adding an **interaction term** between *female* and *married* to the model where *female* and *married* appear separately. This allows the marriage premium to depend on gender, just as it did in equation (7.11). For purposes of comparison, the estimated model with the *female-married* interaction term is

$$\begin{aligned}\widehat{\log(wage)} &= .321 - .110\,female + .231\,married\\ &\qquad (.100)\ \ (.056)\qquad\ (.055)\\ &\quad - .301\,female{\cdot}married + \cdots,\\ &\qquad (.072)\end{aligned} \tag{7.14}$$

where the rest of the regression is necessarily identical to (7.11). Equation (7.14) shows explicitly that there is a statistically significant interaction between gender and marital status. This model also allows us to obtain the estimated wage differential among all four groups, but here we must be careful to plug in the correct combination of zeros and ones.

Setting *female* = 0 and *married* = 0 corresponds to the group single men, which is the base group, as this eliminates *female*, *married*, and *female* · *married*. We can find the intercept for married men by setting *female* = 0 and *married* = 1 in (7.14); this gives an intercept of .321 + .213 = .534, and so on.

Equation (7.14) is just a different way of finding wage differentials across all gender–marital status combinations. It allows us to easily test the null hypothesis that the gender differential does not depend on marital status (equivalently, that the marriage differential does not depend on gender). Equation (7.11) is more convenient for testing for wage differentials between any group and the base group of single men.

EXAMPLE 7.9 **Effects of Computer Usage on Wages**

Krueger (1993) estimates the effects of computer usage on wages. He defines a dummy variable, which we call *compwork*, equal to one if an individual uses a computer at work. Another dummy variable, *comphome*, equals one if the person uses a computer at home. Using 13,379 people from the 1989 Current Population Survey, Krueger (1993, Table 4) obtains

$$\widehat{\log(wage)} = \hat{\beta}_0 + \underset{(.009)}{.177\ compwork} + \underset{(.019)}{.070\ comphome} + \underset{(.023)}{.017\ compwork \cdot comphome} + other\ factors. \qquad [7.15]$$

(The other factors are the standard ones for wage regressions, including education, experience, gender, and marital status; see Krueger's paper for the exact list.) Krueger does not report the intercept because it is not of any importance; all we need to know is that the base group consists of people who do not use a computer at home or at work. It is worth noticing that the estimated return to using a computer at work (but not at home) is about 17.7%. (The more precise estimate is 19.4%.) Similarly, people who use computers at home but not at work have about a 7% wage premium over those who do not use a computer at all. The differential between those who use a computer at both places, relative to those who use a computer in neither place, is about 26.4% (obtained by adding all three coefficients and multiplying by 100), or the more precise estimate 30.2% obtained from equation (7.10).

The interaction term in (7.15) is not statistically significant, nor is it very big economically. But it is causing little harm by being in the equation.

7-4b Allowing for Different Slopes

We have now seen several examples of how to allow different intercepts for any number of groups in a multiple regression model. There are also occasions for interacting dummy variables with explanatory variables that are not dummy variables to allow for **difference in slopes**. Continuing with the wage example, suppose that we wish to test whether the return to education is the same for men and women, allowing for a constant wage differential between men and women (a differential for which we have already found evidence). For simplicity, we include only education and gender in the model. What kind of model allows for different returns to education? Consider the model

$$\log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female)educ + u. \qquad [7.16]$$

If we plug *female* = 0 into (7.16), then we find that the intercept for males is β_0, and the slope on education for males is β_1. For females, we plug in *female* = 1; thus, the intercept for females is $\beta_0 + \delta_0$, and the slope is $\beta_1 + \delta_1$. Therefore, δ_0 measures the difference in intercepts between women and men, and δ_1 measures the difference in the return to education between women and men. Two of the four cases for the signs of δ_0 and δ_1 are presented in Figure 7.2.

Graph (a) shows the case where the intercept for women is below that for men, and the slope of the line is smaller for women than for men. This means that women earn less than men at all levels of education, and the gap increases as *educ* gets larger. In graph (b), the intercept for women is below that for men, but the slope on education is larger for women. This means that women earn less than men at low levels of education, but the gap narrows as education increases. At some point, a woman earns more than a man with the same level of education, and this amount of education is easily found once we have the estimated equation.

How can we estimate model (7.16)? To apply OLS, we must write the model with an interaction between *female* and *educ*:

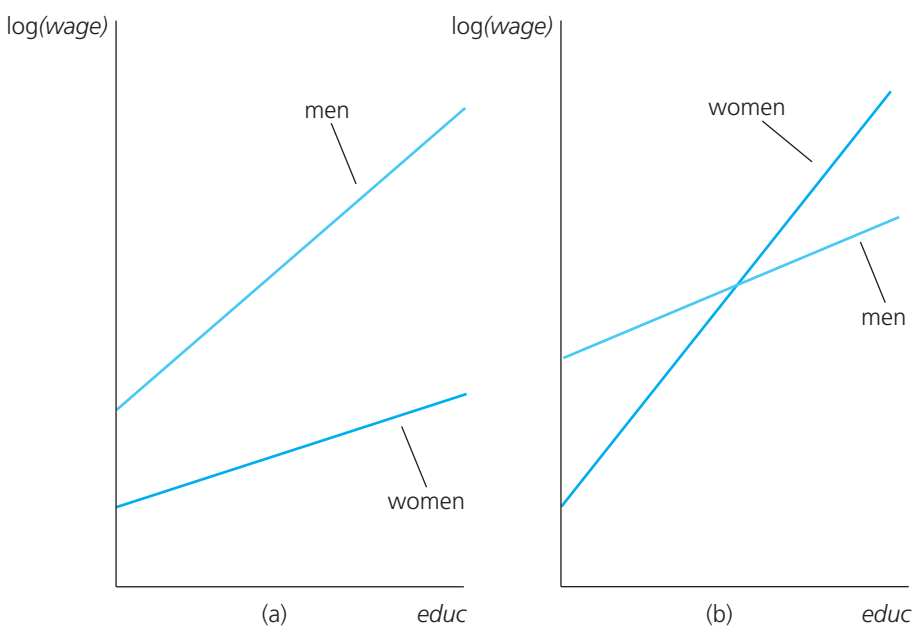
$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u. \qquad [7.17]$$

The parameters can now be estimated from the regression of log(*wage*) on *female*, *educ*, and *female*·*educ*. Obtaining the interaction term is easy in any regression package. Do not be daunted by the odd nature of *female*·*educ*, which is zero for any man in the sample and equal to the level of education for any woman in the sample.

An important hypothesis is that the return to education is the same for women and men. In terms of model (7.17), this is stated as H_0: $\delta_1 = 0$, which means that the slope of log(*wage*) with respect to *educ* is the same for men and women. Note that this hypothesis puts no restrictions on the difference in intercepts, δ_0. A wage differential between men and women is allowed under this null, but it must be the same at all levels of education. This situation is described by Figure 7.1.

We are also interested in the hypothesis that average wages are identical for men and women who have the same levels of education. This means that δ_0 and δ_1 must *both* be zero under the null hypothesis. In equation (7.17), we must use an *F* test to test H_0: $\delta_0 = 0, \delta_1 = 0$. In the model with just an intercept difference, we reject this hypothesis because H_0: $\delta_0 = 0$ is soundly rejected against H_1: $\delta_0 < 0$.

FIGURE 7.2 Graphs of equation (7.16): (a) $\delta_0 < 0, \delta_1 < 0$; (b) $\delta_0 < 0, \delta_1 > 0$.



EXAMPLE 7.10 Log Hourly Wage Equation

We add quadratics in experience and tenure to (7.17):

$$\begin{aligned}\widehat{\log(wage)} &= .389 - .227\,female + .082\,educ \\ &\quad\ (.119)\ \ (.168)\qquad\ (.008) \\ &\quad - .0056\,female{\cdot}educ + .029\,exper - .00058\,exper^2 \\ &\quad\ (.0131)\qquad\qquad (.005)\qquad\ (.00011) \\ &\quad + .032\,tenure - .00059\,tenure^2 \\ &\quad\ (.007)\qquad\ (.00024) \\ n &= 526,\ R^2 = .441.\end{aligned} \tag{7.18}$$

The estimated return to education for men in this equation is .082, or 8.2%. For women, it is .082 − .0056 = .0764, or about 7.6%. The difference, −.56%, or just over one-half a percentage point less for women, is not economically large nor statistically significant: the t statistic is $-.0056/.0131 \approx -.43$. Thus, we conclude that there is no evidence against the hypothesis that the return to education is the same for men and women.

The coefficient on *female*, while remaining economically large, is no longer significant at conventional levels $(t = -1.35)$. Its coefficient and t statistic in the equation without the interaction were −.297 and −8.25, respectively [see equation (7.9)]. Should we now conclude that there is no statistically significant evidence of lower pay for women at the same levels of *educ*, *exper*, and *tenure*? This would be a serious error. Because we have added the interaction *female·educ* to the equation, the coefficient on *female* is now estimated much less precisely than it was in equation (7.9): the standard error has increased by almost fivefold $(.168/.036 \approx 4.67)$. This occurs because *female* and *female·educ* are highly correlated in the sample. In this example, there is a useful way to think about the multicollinearity: in equation (7.17) and the more general equation estimated in (7.18), δ_0 measures the wage differential between women and men when $educ = 0$. Very few people in the sample have very low levels of education, so it is not surprising that we have a difficult time estimating the differential at $educ = 0$ (nor is the differential at zero years of education very informative). More interesting would be to estimate the gender differential at, say, the average education level in the sample (about 12.5). To do this, we would replace *female·educ* with *female*·$(educ - 12.5)$ and rerun the regression; this only changes the coefficient on *female* and its standard error. (See Computer Exercise C7.)

If we compute the F statistic for H_0: $\delta_0 = 0$, $\delta_1 = 0$, we obtain $F = 34.33$, which is a huge value for an F random variable with numerator $df = 2$ and denominator $df = 518$: the p-value is zero to four decimal places. In the end, we prefer model (7.9), which allows for a constant wage differential between women and men.

GOING FURTHER 7.4

How would you augment the model estimated in (7.18) to allow the return to *tenure* to differ by gender?

As a more complicated example involving interactions, we now look at the effects of race and city racial composition on major league baseball player salaries.

EXAMPLE 7.11 Effects of Race on Baseball Player Salaries

Using MLB1, the following equation is estimated for the 330 major league baseball players for which city racial composition statistics are available. The variables *black* and *hispan* are binary indicators for the individual players. (The base group is white players.) The variable *percblck* is the percentage of the team's city that is black, and *perchisp* is the percentage of Hispanics. The other variables

measure aspects of player productivity and longevity. Here, we are interested in race effects after controlling for these other factors.

In addition to including *black* and *hispan* in the equation, we add the interactions *black·percblck* and *hispan·perchisp*. The estimated equation is

$$\begin{aligned}\widehat{\log(salary)} &= 10.34 + .0673\,years + .0089\,gamesyr \\ &\qquad (2.18)\ (.0129)\qquad (.0034) \\ &\quad + .00095\,bavg + .0146\,hrunsyr + .0045\,rbisyr \\ &\qquad (.00151)\qquad (.0164)\qquad (.0076) \\ &\quad + .0072\,runsyr + .0011\,fldperc + .0075\,allstar \\ &\qquad (.0046)\qquad (.0021)\qquad (.0029) \\ &\quad - .198\,black - .190\,hispan + .0125\,black{\cdot}percblck \\ &\qquad (.125)\qquad (.153)\qquad (.0050) \\ &\quad + .0201\,hispan{\cdot}perchisp \\ &\qquad (.0098) \\ n &= 330,\ R^2 = .638. \end{aligned} \tag{7.19}$$

First, we should test whether the four race variables, *black*, *hispan*, *black·percblck*, and *hispan·perchisp*, are jointly significant. Using the same 330 players, the R-squared when the four race variables are dropped is .626. Because there are four restrictions and $df = 330 - 13$ in the unrestricted model, the F statistic is about 2.63, which yields a p-value of .034. Thus, these variables are jointly significant at the 5% level (though not at the 1% level).

How do we interpret the coefficients on the race variables? In the following discussion, all productivity factors are held fixed. First, consider what happens for black players, holding *perchisp* fixed. The coefficient −.198 on *black* literally means that, if a black player is in a city with no blacks ($percblck = 0$), then the black player earns about 19.8% less than a comparable white player. As *percblck* increases—which means the white population decreases, because *perchisp* is held fixed—the salary of blacks increases relative to that for whites. In a city with 10% blacks, log(*salary*) for blacks compared to that for whites is $-.198 + .0125(10) = -.073$, so salary is about 7.3% less for blacks than for whites in such a city. When $percblck = 20$, blacks earn about 5.2% more than whites. The largest percentage of blacks in a city is about 74% (Detroit).

Similarly, Hispanics earn less than whites in cities with a low percentage of Hispanics. But we can easily find the value of *perchisp* that makes the differential between whites and Hispanics equal zero: it must make $-.190 + .0201\,perchisp = 0$, which gives $perchisp \approx 9.45$. For cities in which the percentage of Hispanics is less than 9.45%, Hispanics are predicted to earn less than whites (for a given black population), and the opposite is true if the percentage of Hispanics is above 9.45%. Twelve of the 22 cities represented in the sample have Hispanic populations that are less than 9.45% of the total population. The largest percentage of Hispanics is about 31%.

How do we interpret these findings? We cannot simply claim discrimination exists against blacks and Hispanics, because the estimates imply that whites earn less than blacks and Hispanics in cities heavily populated by minorities. The importance of city composition on salaries might be due to player preferences: perhaps the best black players live disproportionately in cities with more blacks and the best Hispanic players tend to be in cities with more Hispanics. The estimates in (7.19) allow us to determine that some relationship is present, but we cannot distinguish between these two hypotheses.

7-4c Testing for Differences in Regression Functions across Groups

The previous examples illustrate that interacting dummy variables with other independent variables can be a powerful tool. Sometimes, we wish to test the null hypothesis that two populations or groups follow the same regression function, against the alternative that one or more of the slopes differ across the groups. We will also see examples of this in Chapter 13, when we discuss pooling different cross sections over time.

Suppose we want to test whether the same regression model describes college grade point averages for male and female college athletes. The equation is

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

where *sat* is SAT score, *hsperc* is high school rank percentile, and *tothrs* is total hours of college courses. We know that, to allow for an intercept difference, we can include a dummy variable for either males or females. If we want any of the slopes to depend on gender, we simply interact the appropriate variable with, say, *female*, and include it in the equation.

If we are interested in testing whether there is *any* difference between men and women, then we must allow a model where the intercept and all slopes can be different across the two groups:

$$\begin{aligned} cumgpa = {} & \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc \\ & + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u. \end{aligned} \qquad [7.20]$$

The parameter δ_0 is the difference in the intercept between women and men, δ_1 is the slope difference with respect to *sat* between women and men, and so on. The null hypothesis that *cumgpa* follows the same model for males and females is stated as

$$H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0. \qquad [7.21]$$

If one of the δ_j is different from zero, then the model is different for men and women.

Using the spring semester data from the file GPA3, the full model is estimated as

$$\begin{aligned} \widehat{cumgpa} = {} & 1.48 - .353\, female + .0011\, sat + .00075\, female \cdot sat \\ & \quad (0.21)\ (.411)\qquad (.0002)\qquad (.00039) \\ & -.0085\, hsperc - .00055\, female \cdot hsperc + .0023\, tothrs \\ & \quad (.0014)\qquad (.00316)\qquad (.0009) \\ & -.00012\, female \cdot tothrs \\ & \quad (.00163) \\ n = {} & 366, R^2 = .406, \overline{R}^2 = .394. \end{aligned} \qquad [7.22]$$

None of the four terms involving the female dummy variable is very statistically significant; only the *female·sat* interaction has a *t* statistic close to two. But we know better than to rely on the individual *t* statistics for testing a joint hypothesis such as (7.21). To compute the *F* statistic, we must estimate the restricted model, which results from dropping *female* and all of the interactions; this gives an R^2 (the restricted R^2) of about .352, so the *F* statistic is about 8.14; the *p*-value is zero to five decimal places, which causes us to soundly reject (7.21). Thus, men and women athletes do follow different GPA models, even though each term in (7.22) that allows women and men to be different is individually insignificant at the 5% level.

The large standard errors on *female* and the interaction terms make it difficult to tell exactly how men and women differ. We must be very careful in interpreting equation (7.22) because, in obtaining differences between women and men, the interaction terms must be taken into account. If we look only at the *female* variable, we would wrongly conclude that *cumgpa* is about .353 less for women than for men, holding other factors fixed. This is the estimated difference only when *sat*, *hsperc*, and *tothrs* are all set to zero, which is not close to being a possible scenario. At $sat = 1{,}100$, $hsperc = 10$, and $tothrs = 50$, the predicted *difference* between a woman and a man is $-.353 + .00075(1{,}100) - .00055(10) - .00012(50) \approx .461$. That is, the female athlete is predicted to have a GPA that is almost one-half a point higher than the comparable male athlete.

In a model with three variables, *sat*, *hsperc*, and *tothrs*, it is pretty simple to add all of the interactions to test for group differences. In some cases, many more explanatory variables are involved, and then it is convenient to have a different way to compute the statistic. It turns out that the sum of squared residuals form of the F statistic can be computed easily even when many independent variables are involved.

In the general model with k explanatory variables and an intercept, suppose we have two groups; call them $g = 1$ and $g = 2$. We would like to test whether the intercept and all slopes are the same across the two groups. Write the model as

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + \cdots + \beta_{g,k}x_k + u, \qquad [7.23]$$

for $g = 1$ and $g = 2$. The hypothesis that each beta in (7.23) is the same across the two groups involves $k + 1$ restrictions (in the GPA example, $k + 1 = 4$). The unrestricted model, which we can think of as having a group dummy variable and k interaction terms in addition to the intercept and variables themselves, has $n - 2(k + 1)$ degrees of freedom. [In the GPA example, $n - 2(k + 1) = 366 - 2(4) = 358$.] So far, there is nothing new. The key insight is that the sum of squared residuals from the unrestricted model can be obtained from two *separate* regressions, one for each group. Let SSR_1 be the sum of squared residuals obtained estimating (7.23) for the first group; this involves n_1 observations. Let SSR_2 be the sum of squared residuals obtained from estimating the model using the second group (n_2 observations). In the previous example, if group 1 is females, then $n_1 = 90$ and $n_2 = 276$. Now, the sum of squared residuals for the unrestricted model is simply $SSR_{ur} = SSR_1 + SSR_2$. The restricted sum of squared residuals is just the SSR from pooling the groups and estimating a single equation, say SSR_P. Once we have these, we compute the F statistic as usual:

$$F = \frac{[SSR_P - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k + 1)]}{k + 1}, \qquad [7.24]$$

where n is the *total* number of observations. This particular F statistic is usually called the **Chow statistic** in econometrics. Because the Chow test is just an F test, it is only valid under homoskedasticity. In particular, under the null hypothesis, the error variances for the two groups must be equal. As usual, normality is not needed for asymptotic analysis.

To apply the Chow statistic to the GPA example, we need the SSR from the regression that pooled the groups together: this is $SSR_P = 85.515$. The SSR for the 90 women in the sample is $SSR_1 = 19.603$, and the SSR for the men is $SSR_2 = 58.752$. Thus, $SSR_{ur} = 19.603 + 58.752 = 78.355$. The F statistic is $[(85.515 - 78.355)/78.355](358/4) \approx 8.18$; of course, subject to rounding error, this is what we get using the R-squared form of the test in the models with and without the interaction terms. (A word of caution: there is no simple R-squared form of the test if separate regressions have been estimated for each group; the R-squared form of the test can be used only if interactions have been included to create the unrestricted model.)

One important limitation of the traditional Chow test, regardless of the method used to implement it, is that the null hypothesis allows for no differences at all between the groups. In many cases, it is more interesting to allow for an intercept difference between the groups and then to test for slope

differences; we saw one example of this in the wage equation in Example 7.10. There are two ways to allow the intercepts to differ under the null hypothesis. One is to include the group dummy and all interaction terms, as in equation (7.22), but then test joint significance of the interaction terms only. The second approach, which produces an identical statistic, is to form a sum-of-squared-residuals F statistic, as in equation (7.24), but where the restricted SSR, called "SSR_P" in equation (7.24), is obtained using a regression that contains only an intercept shift. Because we are testing k restrictions, rather than $k + 1$, the F statistic becomes

$$F = \frac{[SSR_P - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k + 1)]}{k}.$$

Using this approach in the GPA example, SSR_P is obtained from the regression *cumgpa* on *female*, *sat*, *hsperc*, and *tothrs* using the data for both male and female student athletes.

Because there are relatively few explanatory variables in the GPA example, it is easy to estimate (7.20) and test H_0: $\delta_1 = 0, \delta_2 = 0, \delta_3 = 0$ (with δ_0 unrestricted under the null). The F statistic for the three exclusion restrictions gives a p-value equal to .205, and so we do not reject the null hypothesis at even the 20% significance level.

Failure to reject the hypothesis that the parameters multiplying the interaction terms are all zero suggests that the best model allows for an intercept difference only:

$$\begin{aligned} \widehat{cumgpa} &= 1.39 + .310\,female + .0012\,sat - .0084\,hsperc \\ &\qquad (.18)\ \ (.059)\qquad\quad (.0002)\qquad (.0012) \\ &\quad + .0025\,tothrs \\ &\qquad (.0007) \\ n &= 366,\ R^2 = .398,\ \overline{R}^2 = .392. \end{aligned} \qquad [7.25]$$

The slope coefficients in (7.25) are close to those for the base group (males) in (7.22); dropping the interactions changes very little. However, *female* in (7.25) is highly significant: its t statistic is over 5, and the estimate implies that, at given levels of *sat*, *hsperc*, and *tothrs*, a female athlete has a predicted GPA that is .31 point higher than that of a male athlete. This is a practically important difference.

7-5 A Binary Dependent Variable: The Linear Probability Model

By now, we have learned much about the properties and applicability of the multiple linear regression model. In the last several sections, we studied how, through the use of binary independent variables, we can incorporate qualitative information as explanatory variables in a multiple regression model. In all of the models up until now, the dependent variable y has had *quantitative* meaning (for example, y is a dollar amount, a test score, a percentage, or the logs of these). What happens if we want to use multiple regression to *explain* a qualitative event?

In the simplest case, and one that often arises in practice, the event we would like to explain is a binary outcome. In other words, our dependent variable, y, takes on only two values: zero and one. For example, y can be defined to indicate whether an adult has a high school education; y can indicate whether a college student used illegal drugs during a given school year; or y can indicate whether a firm was taken over by another firm during a given year. In each of these examples, we can let $y = 1$ denote one of the outcomes and $y = 0$ the other outcome.

What does it mean to write down a multiple regression model, such as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \qquad [7.26]$$

when y is a binary variable? Because y can take on only two values, β_j cannot be interpreted as the change in y given a one-unit increase in x_j, holding all other factors fixed: y either changes from zero to one or from one to zero (or does not change). Nevertheless, the β_j still have useful interpretations. If we assume that the zero conditional mean assumption MLR.4 holds, that is, $\mathrm{E}(u|x_1, \ldots, x_k) = 0$, then we have, as always,

$$\mathrm{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

where $\mathbf{x}$ is shorthand for all of the explanatory variables.

The key point is that when y is a binary variable taking on the values zero and one, it is always true that $\mathrm{P}(y = 1|\mathbf{x}) = E(y|\mathbf{x})$: the probability of "success"—that is, the probability that $y = 1$—is the same as the expected value of y. Thus, we have the important equation

$$\mathrm{P}(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \qquad [7.27]$$

which says that the probability of success, say, $p(\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x})$, is a linear function of the x_j. Equation (7.27) is an example of a binary response model, and $\mathrm{P}(y = 1|\mathbf{x})$ is also called the **response probability**. (We will cover other binary response models in Chapter 17.) Because probabilities must sum to one, $\mathrm{P}(y = 0|\mathbf{x}) = 1 - \mathrm{P}(y = 1|\mathbf{x})$ is also a linear function of the x_j.

The multiple linear regression model with a binary dependent variable is called the **linear probability model (LPM)** because the response probability is linear in the parameters β_j, In the LPM, β_j measures the change in the probability of success when x_j changes, holding other factors fixed:

$$\Delta\mathrm{P}(y = 1|\mathbf{x}) = \beta_j \Delta x_j. \qquad [7.28]$$

With this in mind, the multiple regression model can allow us to estimate the effect of various explanatory variables on qualitative events. The mechanics of OLS are the same as before.

If we write the estimated equation as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k,$$

we must now remember that $\hat{y}$ is the predicted probability of success. Therefore, $\hat{\beta}_0$ is the predicted probability of success when each x_j is set to zero, which may or may not be interesting. The slope coefficient $\hat{\beta}_1$ measures the predicted change in the probability of success when x_1 increases by one unit.

To correctly interpret a linear probability model, we must know what constitutes a "success." Thus, it is a good idea to give the dependent variable a name that describes the event $y = 1$. As an example, let *inlf* ("in the labor force") be a binary variable indicating labor force participation by a married woman during 1975: *inlf* = 1 if the woman reports working for a wage outside the home at some point during the year, and zero otherwise. We assume that labor force participation depends on other sources of income, including husband's earnings (*nwifeinc*, measured in thousands of dollars), years of education (*educ*), past years of labor market experience (*exper*), *age*, number of children less than six years old (*kidslt6*), and number of kids between 6 and 18 years of age (*kidsge6*). Using the data in MROZ from Mroz (1987), we estimate the following linear probability model, where 428 of the 753 women in the sample report were in the labor force at some point during 1975:

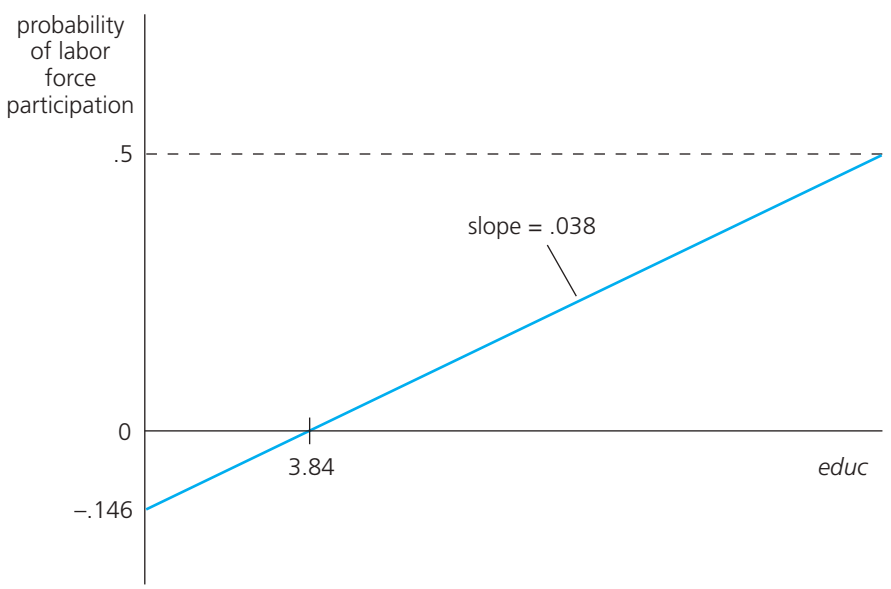
$$\begin{aligned} \widehat{inlf} &= .586 - .0034\, nwifeinc + .038\, educ + .039\, exper \\ &\quad (.154)\ (.0014)\qquad (.007)\qquad (.006) \\ &\quad - .00060\, exper^2 - .016\, age - .262\, kidslt6 + .013\, kidsge6 \\ &\quad (.00018)\qquad (.002)\qquad (.034)\qquad (.013) \\ n &= 753,\ R^2 = .264. \end{aligned} \qquad [7.29]$$

Using the usual t statistics, all variables in (7.29) except *kidsge6* are statistically significant, and all of the significant variables have the effects we would expect based on economic theory (or common sense).

To interpret the estimates, we must remember that a change in the independent variable changes the probability that $inlf = 1$. For example, the coefficient on *educ* means that, everything else in (7.29) held fixed, another year of education increases the probability of labor force participation by .038. If we take this equation literally, 10 more years of education increases the probability of being in the labor force by $.038(10) = .38$, which is a pretty large increase in a probability. The relationship between the probability of labor force participation and *educ* is plotted in Figure 7.3. The other independent variables are fixed at the values $nwifeinc = 50$, $exper = 5$, $age = 30$, $kidslt6 = 1$, and $kidsge6 = 0$ for illustration purposes. The predicted probability is negative until education equals 3.84 years. This should not cause too much concern because, in this sample, no woman has less than five years of education. The largest reported education is 17 years, and this leads to a predicted probability of .5. If we set the other independent variables at different values, the range of predicted probabilities would change. But the marginal effect of another year of education on the probability of labor force participation is always .038.

The coefficient on *nwifeinc* implies that, if $\Delta nwifeinc = 10$ (which means an increase of $10,000), the probability that a woman is in the labor force falls by .034. This is not an especially large effect given that an increase in income of $10,000 is substantial in terms of 1975 dollars. Experience has been entered as a quadratic to allow the effect of past experience to have a diminishing effect on the labor force participation probability. Holding other factors fixed, the estimated change in the probability is approximated as $.039 - 2(.0006)exper = .039 - .0012\, exper$. The point at which past experience has no effect on the probability of labor force participation is $.039/.0012 = 32.5$, which is a high level of experience: only 13 of the 753 women in the sample have more than 32 years of experience.

FIGURE 7.3 Estimated relationship between the probability of being in the labor force and years of education, with other explanatory variables fixed.



Unlike the number of older children, the number of young children has a huge impact on labor force participation. Having one additional child less than six years old reduces the probability of participation by −.262, at given levels of the other variables. In the sample, just under 20% of the women have at least one young child.

This example illustrates how easy linear probability models are to estimate and interpret, but it also highlights some shortcomings of the LPM. First, it is easy to see that, if we plug certain combinations of values for the independent variables into (7.29), we can get predictions either less than zero or greater than one. Because these are predicted probabilities, and probabilities must be between zero and one, this can be a little embarrassing. For example, what would it mean to predict that a woman is in the labor force with a probability of –.10? In fact, of the 753 women in the sample, 16 of the fitted values from (7.29) are less than zero, and 17 of the fitted values are greater than one.

A related problem is that a probability cannot be linearly related to the independent variables for all their possible values. For example, (7.29) predicts that the effect of going from zero children to one young child reduces the probability of working by .262. This is also the predicted drop if the woman goes from having one young child to two. It seems more realistic that the first small child would reduce the probability by a large amount, but subsequent children would have a smaller marginal effect. In fact, when taken to the extreme, (7.29) implies that going from zero to four young children reduces the probability of working by $\Delta\widehat{inlf} = .262(\Delta kidslt6) = .262(4) = 1.048$, which is impossible.

Even with these problems, the linear probability model is useful and often applied in economics. It usually works well for values of the independent variables that are near the averages in the sample. In the labor force participation example, no women in the sample have four young children; in fact, only three women have three young children. Over 96% of the women have either no young children or one small child, and so we should probably restrict attention to this case when interpreting the estimated equation.

Predicted probabilities outside the unit interval are a little troubling when we want to make predictions. Still, there are ways to use the estimated probabilities (even if some are negative or greater than one) to predict a zero-one outcome. As before, let $\hat{y}_i$ denote the fitted values—which may not be bounded between zero and one. Define a predicted value as $\tilde{y}_i = 1$ if $\hat{y}_i \geq .5$ and $\tilde{y}_i = 0$ if $\hat{y}_i < .5$. Now we have a set of predicted values, $\tilde{y}_i$, $i = 1, \ldots, n$, that, like the y_i, are either zero or one. We can use the data on y_I and $\tilde{y}_i$ to obtain the frequencies with which we correctly predict $y_i = 1$ and $y_i = 0$, as well as the proportion of overall correct predictions. The latter measure, when turned into a percentage, is a widely used goodness-of-fit measure for binary dependent variables: the **percent correctly predicted**. An example is given in Computer Exercise C9(v), and further discussion, in the context of more advanced models, can be found in Section 17-1.

Due to the binary nature of y, the linear probability model does violate one of the Gauss-Markov assumptions. When y is a binary variable, its variance, conditional on $\mathbf{x}$, is

$$\text{Var}(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})], \qquad [7.30]$$

where $p(\mathbf{x})$ is shorthand for the probability of success: $p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. This means that, except in the case where the probability does not depend on any of the independent variables, there *must* be heteroskedasticity in a linear probability model. We know from Chapter 3 that this does not cause bias in the OLS estimators of the β_j. But we also know from Chapters 4 and 5 that homoskedasticity is crucial for justifying the usual t and F statistics, even in large samples. Because the standard errors in (7.29) are not generally valid, we should use them with caution. We will show how to correct the standard errors for heteroskedasticity in Chapter 8. It turns out that, in many applications, the usual OLS statistics are not far off, and it is still acceptable in applied work to present a standard OLS analysis of a linear probability model.

EXAMPLE 7.12 A Linear Probability Model of Arrests

Let *arr86* be a binary variable equal to unity if a man was arrested during 1986, and zero otherwise. The population is a group of young men in California born in 1960 or 1961 who have at least one arrest prior to 1986. A linear probability model for describing *arr86* is

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

where

$pcnv$ = the proportion of prior arrests that led to a conviction.
$avgsen$ = the average sentence served from prior convictions (in months).
$tottime$ = months spent in prison since age 18 prior to 1986.
$ptime86$ = months spent in prison in 1986.
$qemp86$ = the number of quarters (0 to 4) that the man was legally employed in 1986.

The data we use are in CRIME1, the same data set used for Example 3.5. Here, we use a binary dependent variable because only 7.2% of the men in the sample were arrested more than once. About 27.7% of the men were arrested at least once during 1986. The estimated equation is

$$\begin{aligned} \widehat{arr86} &= .441 - .162\, pcnv + .0061\, avgsen - .0023\, tottime \\ &\qquad (.017)\ (.021)\qquad (.0065)\qquad (.0050) \\ &\quad - .022\, ptime86 - .043\, qemp86 \\ &\qquad (.005)\qquad\ (.005) \\ n &= 2{,}725,\ R^2 = .0474. \end{aligned} \tag{7.31}$$

The intercept, .441, is the predicted probability of arrest for someone who has not been convicted (and so *pcnv* and *avgsen* are both zero), has spent no time in prison since age 18, spent no time in prison in 1986, and was unemployed during the entire year. The variables *avgsen* and *tottime* are insignificant both individually and jointly (the F test gives $p\text{-value} = .347$), and *avgsen* has a counterintuitive sign if longer sentences are supposed to deter crime. Grogger (1991), using a superset of these data and different econometric methods, found that *tottime* has a statistically significant *positive* effect on arrests and concluded that *tottime* is a measure of human capital built up in criminal activity.

Increasing the probability of conviction does lower the probability of arrest, but we must be careful when interpreting the magnitude of the coefficient. The variable *pcnv* is a proportion between zero and one; thus, changing *pcnv* from zero to one essentially means a change from no chance of being convicted to being convicted with certainty. Even this large change reduces the probability of arrest only by .162; increasing *pcnv* by .5 decreases the probability of arrest by .081.

The incarcerative effect is given by the coefficient on *ptime86*. If a man is in prison, he cannot be arrested. Because *ptime86* is measured in months, six more months in prison reduces the probability of arrest by $.022(6) = .132$. Equation (7.31) gives another example of where the linear probability model cannot be true over all ranges of the independent variables. If a man is in prison all 12 months of 1986, he cannot be arrested in 1986. Setting all other variables equal to zero, the predicted probability of arrest when $ptime86 = 12$ is $.441 - .022(12) = .177$, which is not zero. Nevertheless, if we start from the unconditional probability of arrest, .277, 12 months in prison reduces the probability to essentially zero: $.277 - .022(12) = .013$.

Finally, employment reduces the probability of arrest in a significant way. All other factors fixed, a man employed in all four quarters is .172 less likely to be arrested than a man who is not employed at all.

We can also include dummy independent variables in models with dummy dependent variables. The coefficient measures the predicted difference in probability relative to the base group. For example, if we add two race dummies, *black* and *hispan*, to the arrest equation, we obtain

$$\widehat{arr86} = .380 - .152\,pcnv + .0046\,avgsen - .0026\,tottime$$
$$\quad (.019)\ (.021)\quad (.0064)\quad (.0049)$$
$$- .024\,ptime86 - .038\,qemp86 + .170\,black + .096\,hispan \qquad [7.32]$$
$$\quad (.005)\quad (.005)\quad (.024)\quad (.021)$$
$$n = 2{,}725,\ R^2 = .0682.$$

GOING FURTHER 7.5

What is the predicted probability of arrest for a black man with no prior convictions—so that *pcnv, avgsen, tottime,* and *ptime86* are all zero—who was employed all four quarters in 1986? Does this seem reasonable?

The coefficient on *black* means that, all other factors being equal, a black man has a .17 higher chance of being arrested than a white man (the base group). Another way to say this is that the probability of arrest is 17 percentage points higher for blacks than for whites. The difference is statistically significant as well. Similarly, Hispanic men have a .096 higher chance of being arrested than white men.

7-6 More on Policy Analysis and Program Evaluation

We have seen some examples of models containing dummy variables that can be useful for evaluating policy. Example 7.3 gave an example of program evaluation, where some firms received job training grants and others did not.

As we mentioned earlier, we must be careful when evaluating programs because in most examples in the social sciences the control and treatment groups are not randomly assigned. Consider again the Holzer et al. (1993) study, where we are now interested in the effect of the job training grants on worker productivity (as opposed to amount of job training). The equation of interest is

$$\log(scrap) = \beta_0 + \beta_1 grant + \beta_2 \log(sales) + \beta_3 \log(employ) + u,$$

where *scrap* is the firm's scrap rate, and the latter two variables are included as controls. The binary variable *grant* indicates whether the firm received a grant in 1988 for job training.

Before we look at the estimates, we might be worried that the unobserved factors affecting worker productivity—such as average levels of education, ability, experience, and tenure—might be correlated with whether the firm receives a grant. Holzer et al. point out that grants were given on a first-come, first-served basis. But this is not the same as giving out grants randomly. It might be that firms with less productive workers saw an opportunity to improve productivity and therefore were more diligent in applying for the grants.

Using the data in JTRAIN for 1988—when firms actually were eligible to receive the grants—we obtain

$$\widehat{\log(scrap)} = 4.99 - .052\,grant - .455\log(sales)$$
$$\quad (4.66)\ (.431)\quad (.373)$$
$$+ .639\log(employ) \qquad [7.33]$$
$$\quad (.365)$$
$$n = 50,\ R^2 = .072.$$

(Seventeen out of the fifty firms received a training grant, and the average scrap rate is 3.47 across all firms.) The point estimate of −.052 on *grant* means that, for given *sales* and *employ*, firms receiving a grant have scrap rates about 5.2% lower than firms without grants. This is the direction of the expected effect if the training grants are effective, but the t statistic is very small. Thus, from this cross-sectional analysis, we must conclude that the grants had no effect on firm productivity. We will return to this example in Chapter 9 and show how adding information from a prior year leads to a much different conclusion.

Even in cases where the policy analysis does not involve assigning units to a control group and a treatment group, we must be careful to include factors that might be systematically related to the binary independent variable of interest. A good example of this is testing for racial discrimination. Race is something that is not determined by an individual or by government administrators. In fact, race would appear to be the perfect example of an exogenous explanatory variable, given that it is determined at birth. However, for historical reasons, race is often related to other relevant factors: there are systematic differences in backgrounds across race, and these differences can be important in testing for *current* discrimination.

As an example, consider testing for discrimination in loan approvals. If we can collect data on, say, individual mortgage applications, then we can define the dummy dependent variable *approved* as equal to one if a mortgage application was approved, and zero otherwise. A systematic difference in approval rates across races is an indication of discrimination. However, because approval depends on many other factors, including income, wealth, credit ratings, and a general ability to pay back the loan, we must control for them *if* there are systematic differences in these factors across race. A linear probability model to test for discrimination might look like the following:

$$approved = \beta_0 + \beta_1 nonwhite + \beta_2 income + \beta_3 wealth + \beta_4 credrate + other\ factors.$$

Discrimination against nonwhites is indicated by a rejection of H_0: $\beta_1 = 0$ in favor of H_0: $\beta_1 < 0$, because β_1 is the amount by which the probability of a nonwhite getting an approval differs from the probability of a white getting an approval, given the same levels of other variables in the equation. If *income*, *wealth*, and so on are systematically different across races, then it is important to control for these factors in a multiple regression analysis.

7-6a Program Evaluation and Unrestricted Regression Adjustment

In Section 3-7e, in the context of potential outcomes, we derived an equation that can be used to test the effectiveness of a policy intervention or a program. Letting w again be the binary policy indicator and $x_1, x_2, \ldots, x_k$ the control variables, we obtained the following population regression function:

$$\mathrm{E}(y|w, \mathbf{x}) = \alpha + \tau w + \mathbf{x}\boldsymbol{\gamma} = \alpha + \tau w + \gamma_1 x_1 + \cdots + \gamma_k x_k, \qquad [7.34]$$

where $y = (1 - w)y(0) + wy(1)$ is the observed outcome and $[y(0),y(1)]$ are the potential or counterfactual outcomes.

The reason for including the x_j in (7.34) is to account for the possibility that program participation is not randomly assigned. The problem of participation decisions differing systematically by individual characteristics is often referred to as the **self-selection problem,** with "self being used broadly. For example, children eligible for programs such as Head Start participate largely based on parental decisions. Because family background and structure play a role in Head Start participation decisions, and they also tend to predict child outcomes, we should control for these socioeconomic factors when examining the effects of Head Start [see, for example, Currie and Thomas (1995)].

In the context of causal inference, the assumption that we have sufficient explanatory variables so that, conditional on those variables, program participation is as good as random, is the

unconfoundedness or ignorability assumption introduced in Section 3-7e. When we are mainly interested in estimating the effect of a program or intervention, as indicated by w,the explanatory variables $x_1, x_2, \ldots, x_k$ are often called **covariates**. These are factors that possibly vary with participation decisions as well as with potential outcomes.

The self-selection problem is not restricted to decisions to participate in school or government programs. It is rampant when studying the economic and societal effects of certain behaviors. For example, individuals choose to use illegal drugs or to drink alcohol. If we want to examine the effects of such behaviors on unemployment status, earnings, or criminal behavior, we should be concerned that drug usage might be correlated with factors affecting potential labor or criminal outcomes. Without accounting for systematic differences between those who use drugs and those who do not, we are unlikely to obtain a convincing causal estimate of drug usage.

Self-selection also can be an issue when studying more aggregate units. Cities and states choose whether to implement certain gun control laws, and it is likely that this decision is systematically related to other factors that affect violent crime [see, for example, Kleck and Patterson (1993)]. Hospitals choose to be for profit or nonprofit, and this decision may be related to hospital characteristics that affect patient health outcomes.

Most program evaluations are still based on observational (or nonexperimental) data, and so estimating the simple equation

$$y = \alpha + \tau w + u \qquad [7.35]$$

by OLS is unlikely to produce an unbiased or consistent estimate of the causal effect. By including suitable covariates, estimation of (7.34) is likely to be more convincing. In the context of program evaluation, using the regression

$$y_i \text{ on } w_i, x_{i1}, x_{i2}, \ldots, x_{ik}, i = 1, \ldots, n \qquad [7.36]$$

is a version of what is called **regression adjustment**, and $\hat{\tau}$, the coefficient on w_i is the *regression adjusted estimator.* The idea is that we have used multiple regression with covariates $x_1, x_2, \ldots, x_k$ to adjust for differences across units in estimating the causal effect.

Recall from Section 3.7e that, in addition to unconfoundedness, equation (7.34) was obtained under the strong assumption of a constant treatment effect: $\tau = y_i(1) - y_i(0)$ for all i. We are now in a position to relax this assumption. We still maintain the unconfoundedness, or conditional independence, assumption, which we reproduce here for convenience:

$$w \text{ is independent of } [y(0), y(1)] \text{ conditional on } \mathbf{x} = (x_1, \ldots, x_k) \qquad [7.37]$$

We also assume the conditional means are linear, but now we allow completely separate equations for $y(0)$ and $y(1)$. Written in terms of errors $u(0)$ and $u(1)$,

$$y(0) = \psi_0 + (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\gamma}_0 + u_0 = \psi_0 + \gamma_{0,1}(x_1 - \eta_1) + \cdots + \gamma_{0,k}(x_k - \eta_k) + u(0) \qquad [7.38]$$

$$y(1) = \psi_1 + (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\gamma}_1 + u_1 = \psi_1 + \gamma_{1,1}(x_1 - \eta_1) + \cdots + \gamma_{1,k}(x_k - \eta_k) + u(1) \qquad [7.39]$$

where $\eta_j = \mathrm{E}(x_j)$ is the population mean of x_j, $\psi_0 = \mathrm{E}[y(0)]$, and $\psi_1 = \mathrm{E}[y(1)]$. The covariates x_j have been centered about their means so that the intercepts, ψ_0 and ψ_1, are the expected values of the two potential outcomes. Equations (7.38) and (7.39) allow the treatment effect for unit i, $y_i(1) - y_i(0)$, to depend on observables $\mathbf{x}_i$ and the unobservables. For unit i, the treatment effect is

$$te_i = y_i(1) - y_i(0) = (\psi_1 - \psi_0) + (\mathbf{x}_i - \boldsymbol{\eta})(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0) + [u_i(1) - u_i(0)]$$

The average treatment effect, which we call τ in this section, is $\tau = \psi_1 - \psi_0$ because

$$\begin{aligned} \mathrm{E}(te_i) &= (\psi_1 - \psi_0) + \mathrm{E}\{\mathbf{x}_i - \boldsymbol{\eta})(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0) + [u_i(1) - u_i(0)]\} \\ &= \tau + \mathbf{0} \cdot (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0) + 0 = \tau, \end{aligned}$$

where $(\mathbf{x}_i - \boldsymbol{\eta})$ has a zero mean by construction and $u_i(0)$, $u_i(1)$ has zero mean because they are the errors obtained from conditional expectations. The observed outcome $y_i = y_i(0) + w_i[y_i(1) - y_i(0)]$ can be written as

$$y_i = \psi_0 + \tau w_i + (\mathbf{x}_i - \boldsymbol{\eta})\boldsymbol{\gamma}_0 + w_i(\mathbf{x}_i - \boldsymbol{\eta})\boldsymbol{\delta} + u_i(0) + w_i[u_i(1) - u_i(0)] \quad [7.40]$$

where $\boldsymbol{\delta} = (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0)$. If we define

$$u_i = u_i(0) + w_i[u_i(1) - u_i(0)]$$

the unconfoundedness implies

$$\begin{aligned} \mathrm{E}(u_i|w_i,\mathbf{x}_i) &= \mathrm{E}[u_i(0)|w_i,\mathbf{x}_i] + w_i\mathrm{E}\{[u_i(1) - u_i(0)]|w_i,\mathbf{x}_i\} \\ &= \mathrm{E}[u_i(0)|\mathbf{x}_i] + w_i\mathrm{E}\{[u_i(1) - u_i(0)]|\mathbf{x}_i\} = 0. \end{aligned} \quad [7.41]$$

Equations (7.40) and (7.41) suggest that we run a regression that includes a full set of interactions between w_i and demeaned controls. To implement this method, we also need to replace the unknown population means, η_j, with the sample averages (across the entire sample of n observations), $\bar{x}_j$. This leads to the regression

$$y_i \text{ on } w_i, x_{i1}, \ldots, x_{ik}, w_i \cdot (x_{i1} - \bar{x}_1), \ldots, w_i \cdot (x_{ik} - \bar{x}_k) \quad [7.42]$$

using all n observations. The coefficient on w_i, $\hat{\tau}$, is the average causal effect or average treatment effect. We can determine how the treatment effect varies with the x_j by multiplying $(x_j - \bar{x}_j)$ by the coefficient on the interaction term, $\hat{\delta}_j$. Note that we do not have to demean the x_j when they appear by themselves, as failing to do so only changes the intercept in the regression. But it is critical to demean the x_j before constructing the interactions in order to obtain the average treatment effect as the coefficient on w_i.

The estimateof τ from (7.42) will be different than that from (7.36), as (7.36) omits the k interaction terms. In the literature, the phrase "regression adjustment" often refers to the more flexible regression in (7.42). For emphasis, one can use the terms *restricted regression adjustment* (RRA) (and use the notation $\hat{\tau}_{rra}$) and *unrestricted regression adjustment* (URA) (using $\hat{\tau}_{ura}$) for (7.36) and (7.42), respectively.

It turns out that the estimate $\hat{\tau}$ from (7.42) can be obtained from two separate regressions, just as when computing the Chow statistic from Section 7.4c. Working through the details is informative, as it emphasizes the counterfactual nature of the unrestricted regression adjustment. First, we run regressions separately for the control and treatment groups. For the control group, we use the n_0 observations with $w_i = 0$ and run the regression

$$y_i \text{ on } x_{i1}, x_{i2}, \ldots, x_{ik}$$

and obtain the interercept $\hat{\alpha}_0$ and k slope estimates $\hat{\gamma}_{0,1}, \hat{\gamma}_{0,2}, \ldots, \hat{\gamma}_{0,k}$. We do the same using the n_1 observations with $w_i = 1$, and obtain the intercept $\hat{\alpha}_1$ and the slopes $\hat{\gamma}_{1,1}, \hat{\gamma}_{1,2}, \ldots, \hat{\gamma}_{1,k}$.

Now here is where we use counterfactual reasoning: for every unit i in the sample, we predict $y_i(0)$ and $y_i(1)$ regardless of whether the unit was in the control or treatment group. Define the predicted values as

$$\begin{aligned} \hat{y}_i^{(0)} &= \hat{\alpha}_0 + \hat{\gamma}_{0,1}x_{i1} + \hat{\gamma}_{0,2}x_{i2} + \cdots + \hat{\gamma}_{0,k}x_{ik} \\ \hat{y}_i^{(1)} &= \hat{\alpha}_1 + \hat{\gamma}_{1,1}x_{i1} + \hat{\gamma}_{1,2}x_{i2} + \cdots + \hat{\gamma}_{1,k}x_{ik} \end{aligned}$$

for all i. In other words, we plug the explanatory variables for unit i into *both* regression functions to predict the outcomes in the two states of the world: the control state and the treated state. It is then natural to estimate the average treatment effect as

$$n^{-1}\sum_{i=1}^{n}[\hat{y}_i^{(1)} - \hat{y}_i^{(0)}] = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\gamma}_{1,1} - \hat{\gamma}_{0,1})\bar{x}_1 + \cdots + (\hat{\gamma}_{1,k} - \hat{\gamma}_{0,k})\bar{x}_k. \quad [7.43]$$

With some algebra, one can show that (7.43) produces the $\hat{\tau}$ from regression (7.40). Thus, two seemingly different ways of using multiple regression to adjust for differences in units lead to the same estimate of the ATE.

Most regression packages are designed to make calculation of (7.43) easy, because they compute a predicted value for all observations that have information on the $\{x_{ij}: j = 1, \ldots, k\}$ whether or not unit i was used in the estimation. However, obtaining a proper standard error when computing (7.43) "by hand" can be tricky, although some econometrics software has the calculation built in. Regression (7.42) always produces a valid standard error of $\hat{\tau}$ but requires obtaining the k interaction terms after demeaning each x_j. Incidentally, we can obtain the Chow test that allows different intercepts by testing whether the k interaction terms are jointly significant using an F statistic. If we fail to reject, we could return to the regression (7.36) that imposes common slopes.

EXAMPLE 7.13 **Evaluating a Job Training Program using Unrestricted Regression Adjustment**

The data in JTRAIN98 were used in Examples 3.11 and 4.11 to estimate the effects of a job training program. The variable we would like to explain, $y = earn98$, is labor market earnings in 1998, the year following the job training program (which took place in 1997). The earnings variable is measured in thousands of dollars. The variable $w = train$ is the binary participation (or "treatment") indicator. We use the same controls as in Example 4.11—*earn*96, *educ, age,* and *married*—but now we use unrestricted regression adjustment. For comparison purposes, the simple difference-in-means estimate is $\tau_{diffmeams} = -2.05$ (se = 0.48) and the restricted regression adjusted estimate, reported in equation (4.52), is $\hat{\tau}_{rra} = 2.44$ (se = 0.44). The estimated equation with full interactions is

$$\begin{aligned}\widehat{earn98} = {} & 5.08 + 3.11\, train + .353\, earn96 + .378\, educ - .196\, age + 2.76\, married \\ & \quad (1.39)\ (0.53) \qquad (.020) \qquad (.078) \qquad (.023) \qquad (0.55) \\ & + .133\, train \cdot (earn96 - \overline{earn96}) - .035\, train \cdot (educ - \overline{educ}) \\ & \quad (0.054) \qquad\qquad\qquad (.137) \\ & + .058\, train \cdot (age - \overline{age}) - .993\, train \cdot (married - \overline{married}) \\ & \quad (.041) \qquad\qquad\qquad (.883)\end{aligned} \tag{7.44}$$

$$n = 1,130,\ R^2 = 0.409.$$

The estimated average treatment effect is the coefficient on *train*, $\hat{\tau}_{ura} = 3.11$ (se = 0.53), which is very statistically significant with $t_{train} > 5.8$. It is also notably higher than the restricted RA estimate, although the F statistic for joint significance of the interaction terms gives p-value ≈ 0.113, and so the interaction terms are not jointly significant at the 10% level.

Example 7.13 warrants some final comments. First, to obtain the average treatment effect as the coefficient on *train*, all explanatory variables, including the dummy variable *married*, must be demeaned before creating the interaction term with train. Using *train* · *married* in place of the final interaction forces the coefficient on *train* to be the average treatment effect for unmarried men—where the averages are across *educ*96, *educ*, and *age*. The estimate turns out to be 3.79 (se = 0.81). The coefficient on *train* · *married* would be unchanged from that in (7.44), −. 993 (se = 0.883), and is still interpreted as the difference in the ATE between married and unmarried men.

In using regression adjustment to estimate the effects of something like a job training program, there is always the possibility that our control variables, $x_1, x_2, \ldots, x_k$, are not sufficient for fully overcoming self-selection into participation. One must be on guard at all times unless w is known to have been randomized. With observational data, the possibilityof finding a spurious effect—in either

direction—is often quite high, even with a rich set of x_j. A good example of this is contained in Currie and Cole (1993). These authors examine the effect of AFDC (Aid to Families with Dependent Children) participation on the birth weight of a child. Even after controlling for a variety of family and background characteristics, the authors obtain OLS estimates that imply participation in AFDC lowers birth weight. As the authors point out, it is hard to believe that AFDC participation itself causes lower birth weight. [See Currie (1995) for additional examples.]

Using a different econometric method that we discuss in Chapter 15, Currie and Cole find evidence for either no effect or a positive effect of AFDC participation on birth weight. When the self-selection problem causes standard multiple regression analysis to be biased due to a lack of sufficient control variables, the more advanced methods covered in Chapters 13,14, and 15, can be used instead.

7-7 Interpreting Regression Results with Discrete Dependent Variables

A binary response is the most extreme form of a discrete random variable: it takes on only two values, zero and one. As we discussed in Section 7-5, the parameters in a linear probability model can be interpreted as measuring the change in the *probability* that $y = 1$ due to a one-unit increase in an explanatory variable. We also discussed that, because y is a zero-one outcome, $\mathrm{P}(y = 1) = \mathrm{E}(y)$, and this equality continues to hold when we condition on explanatory variables.

Other discrete dependent variables arise in practice, and we have already seen some examples, such as the number of times someone is arrested in a given year (Example 3.5). Studies on factors affecting fertility often use the number of living children as the dependent variable in a regression analysis. As with number of arrests, the number of living children takes on a small set of integer values, and zero is a common value. The data in FERTIL2, which contains information on a large sample of women in Botswana is one such example. Often demographers are interested in the effects of education on fertility, with special attention to trying to determine whether education has a causal effect on fertility. Such examples raise a question about how one interprets regression coefficients: after all, one cannot have a fraction of a child.

To illustrate the issues, the regression below uses the data in FERTIL2:

$$\begin{aligned} \widehat{children} &= -1.997 + .175\, age - .090\, educ \\ &\qquad (.094)\ \ (.003)\qquad (.006) \\ n &= 4{,}361,\ R^2 = .560. \end{aligned} \tag{7.45}$$

At this time, we ignore the issue of whether this regression adequately controls for all factors that affect fertility. Instead we focus on interpreting the regression coefficients.

Consider the main coefficient of interest, $\hat{\beta}_{educ} = -.090$. If we take this estimate literally, it says that each additional year of education reduces the estimated number of children by .090—something obviously impossible for any particular woman. A similar problem arises when trying to interpret $\hat{\beta}_{age} = .175$. How can we make sense of these coefficients?

To interpret regression results generally, even in cases where y is discrete and takes on a small number of values, it is useful to remember the interpretation of OLS as estimating the effects of the x_j on the *expected* (or *average*) value of y. Generally, under Assumptions MLR.1 and MLR.4,

$$\mathrm{E}(y|x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{7.46}$$

Therefore, β_j is the effect of a ceteris paribus increase of x_j on the expected value of y. As we discussed in Section 6-4, for a given set of x_j values we interpret the predicted value, $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$, as an estimate of $\mathrm{E}(y|x_1, x_2, \ldots, x_k)$. Therefore, $\hat{\beta}_j$ is our estimate of how the *average* of y changes when $\Delta x_j = 1$ (keeping other factors fixed).

Seen in this light, we can now provide meaning to regression results as in equation (7.45). The coefficient $\hat{\beta}_{educ} = -.090$ means that we estimate that *average* fertility falls by .09 children given one more year of education. A nice way to summarize this interpretation is that if each woman in a group of 100 obtains another year of education, we estimate there will be nine fewer children among them.

Adding dummy variables to regressions when y is itself discrete causes no problems when we interpret the estimated effect in terms of average values. Using the data in FERTIL2 we get

$$\widehat{children} = -2.071 + .177\,age - .079\,educ - .362\,electric \qquad [7.47]$$
$$(.095)\ (.003)\qquad (.006)\qquad (.068)$$
$$n = 4{,}358,\ R^2 = .562,$$

where *electric* is a dummy variable equal to one if the woman lives in a home with electricity. Of course it cannot be true that a particular woman who has electricity has .362 less children than an otherwise comparable woman who does not. But we can say that when comparing 100 women with electricity to 100 women without—at the same age and level of education—we estimate the former group to have about 36 fewer children.

Incidentally, when y is discrete the linear model does not always provide the best estimates of partial effects on $\mathrm{E}(y|x_1, x_2, \ldots, x_k)$. Chapter 17 contains more advanced models and estimation methods that tend to fit the data better when the range of y is limited in some substantive way. Nevertheless, a linear model estimated by OLS often provides a good approximation to the true partial effects, at least on average.

Summary

In this chapter, we have learned how to use qualitative information in regression analysis. In the simplest case, a dummy variable is defined to distinguish between two groups, and the coefficient estimate on the dummy variable estimates the ceteris paribus difference between the two groups. Allowing for more than two groups is accomplished by defining a set of dummy variables: if there are g groups, then $g - 1$ dummy variables are included in the model. All estimates on the dummy variables are interpreted relative to the base or benchmark group (the group for which no dummy variable is included in the model).

Dummy variables are also useful for incorporating ordinal information, such as a credit or a beauty rating, in regression models. We simply define a set of dummy variables representing different outcomes of the ordinal variable, allowing one of the categories to be the base group.

Dummy variables can be interacted with quantitative variables to allow slope differences across different groups. In the extreme case, we can allow each group to have its own slope on every variable, as well as its own intercept. The Chow test can be used to detect whether there are any differences across groups. In many cases, it is more interesting to test whether, after allowing for an intercept difference, the slopes for two different groups are the same. A standard F test can be used for this purpose in an unrestricted model that includes interactions between the group dummy and all variables.

The linear probability model, which is simply estimated by OLS, allows us to explain a binary response using regression analysis. The OLS estimates are now interpreted as changes in the probability of "success" $(y = 1)$, given a one-unit increase in the corresponding explanatory variable. The LPM does have some drawbacks: it can produce predicted probabilities that are less than zero or greater than one, it implies a constant marginal effect of each explanatory variable that appears in its original form, and it contains heteroskedasticity. The first two problems are often not serious when we are obtaining estimates

of the partial effects of the explanatory variables for the middle ranges of the data. Heteroskedasticity does invalidate the usual OLS standard errors and test statistics, but, as we will see in the next chapter, this is easily fixed in large enough samples.

Section 7.6 provides a discussion of how binary variables are used to evaluate policies and programs. As in all regression analysis, we must remember that program participation, or some other binary regressor with policy implications, might be correlated with unobserved factors that affect the dependent variable, resulting in the usual omitted variables bias.

We ended this chapter with a general discussion of how to interpret regression equations when the dependent variable is discrete. The key is to remember that the coefficients can be interpreted as the effects on the expected value of the dependent variable.

Key Terms

Base Group
Benchmark Group
Binary Variable
Chow Statistic
Control Group
Covariates
Difference in Slopes
Dummy Variable Trap
Dummy Variables
Experimental Group
Interaction Term
Intercept Shift
Linear Probability Model (LPM)
Ordinal Variable
Percent Correctly Predicted
Policy Analysis
Program Evaluation
Regression adjustment
Response Probability
Self-Selection Problem
Treatment Group
Uncentered *R*-Squared
Zero-One Variable

Problems

1 Using the data in SLEEP75 (see also Problem 3 in Chapter 3), we obtain the estimated equation

$$\begin{aligned}\widehat{sleep} &= 3{,}840.83 - .163\,totwrk - 11.71\,educ - 8.70\,age \\ &\qquad (235.11)\ (.018)\qquad\quad (5.86)\qquad (11.21) \\ &\quad + .128\,age^2 + 87.75\,male \\ &\qquad (.134)\qquad (34.33) \\ n &= 706,\ R^2 = .123,\ \overline{R}^2 = .117.\end{aligned}$$

The variable *sleep* is total minutes per week spent sleeping at night, *totwrk* is total weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

(i) All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?

(ii) Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

(iii) What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

2 The following equations were estimated using the data in BWGHT:

$$\begin{aligned}\widehat{\log(bwght)} &= 4.66 - .0044\,cigs + .0093\log(faminc) + .016\,parity \\ &\qquad (.22)\ (.0009)\qquad (.0059)\qquad\quad (.006) \\ &\quad + .027\,male + .055\,white \\ &\qquad (.010)\qquad (.013) \\ n &= 1{,}388,\ R^2 = .0472\end{aligned}$$

and

$$\widehat{\log(bwght)} = 4.65 - .0052\,cigs + .0110\log(faminc) + .017\,parity$$
$$(.38)\ (.0010)\qquad (.0085)\qquad (.006)$$
$$+ .034\,male + .045\,white - .0030\,motheduc + .0032\,fatheduc$$
$$(.011)\qquad (.015)\qquad (.0030)\qquad (.0026)$$
$$n = 1{,}191,\ R^2 = .0493.$$

The variables are defined as in Example 4.9, but we have added a dummy variable for whether the child is male and a dummy variable indicating whether the child is classified as white.

(i) In the first equation, interpret the coefficient on the variable *cigs*. In particular, what is the effect on birth weight from smoking 10 more cigarettes per day?

(ii) How much more is a white child predicted to weigh than a nonwhite child, holding the other factors in the first equation fixed? Is the difference statistically significant?

(iii) Comment on the estimated effect and statistical significance of *motheduc*.

(iv) From the given information, why are you unable to compute the F statistic for joint significance of *motheduc* and *fatheduc*? What would you have to do to compute the F statistic?

3 Using the data in GPA2, the following equation was estimated:

$$\widehat{sat} = 1{,}028.10 + 19.30\,hsize - 2.19\,hsize^2 - 45.09\,female$$
$$(6.29)\ (3.83)\qquad (.53)\qquad (4.29)$$
$$- 169.81\,black + 62.31\,female{\cdot}black$$
$$(12.71)\qquad (18.15)$$
$$n = 4{,}137,\ R^2 = .0858.$$

The variable *sat* is the combined SAT score; *hsize* is size of the student's high school graduating class, in hundreds; *female* is a gender dummy variable; and *black* is a race dummy variable equal to one for blacks, and zero otherwise.

(i) Is there strong evidence that $hsize^2$ should be included in the model? From this equation, what is the optimal high school size?

(ii) Holding *hsize* fixed, what is the estimated difference in SAT score between nonblack females and nonblack males? How statistically significant is this estimated difference?

(iii) What is the estimated difference in SAT score between nonblack males and black males? Test the null hypothesis that there is no difference between their scores, against the alternative that there is a difference.

(iv) What is the estimated difference in SAT score between black females and nonblack females? What would you need to do to test whether the difference is statistically significant?

4 An equation explaining chief executive officer salary is

$$\widehat{\log(salary)} = 4.59 + .257\log(sales) + .011\,roe + .158\,finance$$
$$(.30)\ (.032)\qquad (.004)\qquad (.089)$$
$$+ .181\,consprod - .283\,utility$$
$$(.085)\qquad (.099)$$
$$n = 209,\ R^2 = .357.$$

The data used are in CEOSAL1, where *finance*, *consprod*, and *utility* are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

(i) Compute the approximate percentage difference in estimated salary between the utility and transportation industries, holding *sales* and *roe* fixed. Is the difference statistically significant at the 1% level?

(ii) Use equation (7.10) to obtain the exact percentage difference in estimated salary between the utility and transportation industries and compare this with the answer obtained in part (i).

(iii) What is the approximate percentage difference in estimated salary between the consumer products and finance industries? Write an equation that would allow you to test whether the difference is statistically significant.

5 In Example 7.2, let *noPC* be a dummy variable equal to one if the student does not own a PC, and zero otherwise.

(i) If *noPC* is used in place of *PC* in equation (7.6), what happens to the intercept in the estimated equation? What will be the coefficient on *noPC*? (*Hint:* Write $PC = 1 - noPC$ and plug this into the equation $\widehat{colGPA} = \hat{\beta}_0 + \hat{\delta}_0 PC + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT$.)

(ii) What will happen to the *R*-squared if *noPC* is used in place of *PC*?

(iii) Should *PC* and *noPC* both be included as independent variables in the model? Explain.

6 To test the effectiveness of a job training program on the subsequent wages of workers, we specify the model

$$\log(wage) = \beta_0 + \beta_1 train + \beta_2 educ + \beta_3 exper + u,$$

where *train* is a binary variable equal to unity if a worker participated in the program. Think of the error term u as containing unobserved worker ability. If less able workers have a greater chance of being selected for the program, and you use an OLS analysis, what can you say about the likely bias in the OLS estimator of β_1? (*Hint:* Refer back to Chapter 3.)

7 In the example in equation (7.29), suppose that we define *outlf* to be one if the woman is out of the labor force, and zero otherwise.

(i) If we regress *outlf* on all of the independent variables in equation (7.29), what will happen to the intercept and slope estimates? (*Hint:* $inlf = 1 - outlf$. Plug this into the population equation $inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \cdots$ and rearrange.)

(ii) What will happen to the standard errors on the intercept and slope estimates?

(iii) What will happen to the *R*-squared?

8 Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: "On how many separate occasions last month did you smoke marijuana?"

(i) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by $x\%$."

(ii) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

(iii) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.

(iv) Using the model in part (iii), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

(v) What are some potential problems with drawing causal inference using the survey data that you collected?

9 Let d be a dummy (binary) variable and let z be a quantitative variable. Consider the model

$$y = \beta_0 + \delta_0 d + \beta_1 z + \delta_1 d \cdot z + u;$$

this is a general version of a model with an interaction between a dummy variable and a quantitative variable. [An example is in equation (7.17).]

(i) Because it changes nothing important, set the error to zero, $u = 0$. Then, when $d = 0$ we can write the relationship between y and z as the function $f_0(z) = \beta_0 + \beta_1 z$. Write the same relationship when $d = 1$, where you should use $f_1(z)$ on the left-hand side to denote the linear function of z.

(ii) Assuming that $\delta_1 \neq 0$ (which means the two lines are not parallel), show that the value of z^* such that $f_0(z^*) = f_1(z^*)$ is $z^* = -\delta_0/\delta_1$. This is the point at which the two lines intersect [as in Figure 7.2 (b)]. Argue that z^* is positive if and only if δ_0 and δ_1 have opposite signs.

(iii) Using the data in TWOYEAR, the following equation can be estimated:

$$\widehat{\log(wage)} = 2.289 - .357\, female + .50\, totcoll + .030\, female \cdot totcoll$$
$$(0.011)\ (.015)\qquad (.003)\qquad (.005)$$
$$n = 6{,}763,\ R^2 = .202,$$

where all coefficients and standard errors have been rounded to three decimal places. Using this equation, find the value of *totcoll* such that the predicted values of log(*wage*) are the same for men and women.

(iv) Based on the equation in part (iii), can women realistically get enough years of college so that their earnings catch up to those of men? Explain.

10 For a child i living in a particular school district, let $voucher_i$ be a dummy variable equal to one if a child is selected to participate in a school voucher program, and let $score_i$ be that child's score on a subsequent standardized exam. Suppose that the participation variable, $voucher_i$, is completely randomized in the sense that it is independent of both observed and unobserved factors that can affect the test score.

(i) If you run a simple regression $score_i$ on $voucher_i$ using a random sample of size n, does the OLS estimator provide an unbiased estimator of the effect of the voucher program?

(ii) Suppose you can collect additional background information, such as family income, family structure (e.g., whether the child lives with both parents), and parents' education levels. Do you need to control for these factors to obtain an unbiased estimator of the effects of the voucher program? Explain.

(iii) Why should you include the family background variables in the regression? Is there a situation in which you would not include the background variables?

11 The following equations were estimated using the data in ECONMATH, with standard errors reported under coefficients. The average class score, measured as a percentage, is about 72.2; exactly 50% of the students are male; and the average of *colgpa* (grade point average at the start of the term) is about 2.81.

$$\widehat{score} = 32.31 + 14.32\, colgpa$$
$$(2.00)\quad (0.70)$$
$$n = 856,\ R^2 = .329,\ \overline{R}^2 = .328.$$

$$\widehat{score} = 29.66 + 3.83\, male + 14.57\, colgpa$$
$$(2.04)\ (0.74)\qquad (0.69)$$
$$n = 856,\ R^2 = .349,\ \overline{R}^2 = .348.$$

$$\widehat{score} = 30.36 + 2.47\, male + 14.33\, colgpa + 0.479\, male \cdot colgpa$$
$$(2.86)\ (3.96)\qquad (0.98)\qquad (1.383)$$
$$n = 856,\ R^2 = .349,\ \overline{R}^2 = .347.$$

$$\widehat{score} = 30.36 + 3.82\, male + 14.33\, colgpa + 0.479\, male \cdot (colgpa - 2.81)$$
$$(2.86)\ (0.74)\qquad (0.98)\qquad (1.383)$$
$$n = 856,\ R^2 = .349,\ \overline{R}^2 = .347.$$

(i) Interpret the coefficient on *male* in the second equation and construct a 95% confidence interval for β_{male}. Does the confidence interval exclude zero?

(ii) In the second equation, why is the estimate on *male* so imprecise? Should we now conclude that there are no gender differences in *score* after controlling for *colgpa*? [*Hint*: You might want to compute an F statistic for the null hypothesis that there is no gender difference in the model with the interaction.]

(iii) Compared with the third equation, why is the coefficient on *male* in the last equation so much closer to that in the second equation and just as precisely estimated?

12 Consider Example 7.11, where, prior to computing the interaction between the race/ethnicity of a player and the city's racial composition, we center the city composition variables about the sample averages, $\overline{percblck}$ and $\overline{perchisp}$ (which are, approximately, 16.55 and 10.82, respectively). The resulting estimated equation is

$$\widehat{\log(salary)} = 10.23 + .0673\, years + .0089\, gamesyr + .00095\, bavg + .0146\, hrunsyr$$
$$(2.18)\ (.0129)\qquad (.0034)\qquad (.00151)\qquad (.0164)$$
$$+ .0045\, rbisyr + .0072\, runsyr + .0011\, fldperc + .0075\, allstar$$
$$(.0076)\qquad (0.0046)\qquad (.0021)\qquad (.0029)$$
$$+ .0080\, black + .0273\, hispan + .0125\, black \cdot (percblck - \overline{percblck})$$
$$(.0840)\qquad (.1084)\qquad (.0050)$$
$$+ .0201\, hispan \cdot (perchisp - \overline{perchisp})$$
$$(.0098)$$
$$n = 330,\ R^2 = 0.638.$$

(i) Why are the coefficients on *black* and *hispan* now so much different than those reported in equation (7.19)? In particular, how can you interpret these coefficients?

(ii) What do you make of the fact that neither *black* nor *hispan* is statistically significant in the above equation?

(iii) In comparing the above equation to (7.19), has anything else changed? Why or why not?

13 (i) In the context of potential outcomes with a sample of size n, let $[y_i(0), y_i(1)]$ denote the pair of potential outcomes for unit i. Define the averages

$$\overline{y(0)} = n^{-1}\sum_{i=1}^{n} y_i(0)$$
$$\overline{y(1)} = n^{-1}\sum_{i=1}^{n} y_i(1)$$

and define the *sample average treatment effect* (SATE) as $SATE = \overline{y(1)} - \overline{y(0)}$. Can you compute the SATE given the typical program evaluation data set?

(ii) Let $\overline{y}_0$ and $\overline{y}_1$ be the sample averages of the observed y_i for the control and treated groups, respectively. Show howthese differ from $\overline{y(0)}$ and $\overline{y(1)}$.

Computer Exercises

C1 Use the data in GPA1 for this exercise.

(i) Add the variables *mothcoll* and *fathcoll* to the equation estimated in (7.6) and report the results in the usual form. What happens to the estimated effect of PC ownership? Is *PC* still statistically significant?

(ii) Test for joint significance of *mothcoll* and *fathcoll* in the equation from part (i) and be sure to report the *p*-value.

(iii) Add $hsGPA^2$ to the model from part (i) and decide whether this generalization is needed.

C2 Use the data in WAGE2 for this exercise.

(i) Estimate the model

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

and report the results in the usual form. Holding other factors fixed, what is the approximate difference in monthly salary between blacks and nonblacks? Is this difference statistically significant?

(ii) Add the variables $exper^2$ and $tenure^2$ to the equation and show that they are jointly insignificant at even the 20% level.

(iii) Extend the original model to allow the return to education to depend on race and test whether the return to education does depend on race.

(iv) Again, start with the original model, but now allow wages to differ across four groups of people: married and black, married and nonblack, single and black, and single and nonblack. What is the estimated wage differential between married blacks and married nonblacks?

C3 A model that allows major league baseball player salary to differ by position is

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + \beta_6 runsyr + \beta_7 fldperc + \beta_8 allstar + \beta_9 frstbase + \beta_{10} scndbase + \beta_{11} thrdbase + \beta_{12} shrtstop + \beta_{13} catcher + u,$$

where outfield is the base group.

(i) State the null hypothesis that, controlling for other factors, catchers and outfielders earn, on average, the same amount. Test this hypothesis using the data in MLB1 and comment on the size of the estimated salary differential.

(ii) State and test the null hypothesis that there is no difference in average salary across positions, once other factors have been controlled for.

(iii) Are the results from parts (i) and (ii) consistent? If not, explain what is happening.

C4 Use the data in GPA2 for this exercise.

(i) Consider the equation

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

where *colgpa* is cumulative college grade point average; *hsize* is size of high school graduating class, in hundreds; *hsperc* is academic percentile in graduating class; *sat* is combined SAT score; *female* is a binary gender variable; and *athlete* is a binary variable, which is one for student athletes. What are your expectations for the coefficients in this equation? Which ones are you unsure about?

(ii) Estimate the equation in part (i) and report the results in the usual form. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?

(iii) Drop *sat* from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part (ii).

(iv) In the model from part (i), allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no ceteris paribus difference between women athletes and women nonathletes.

(v) Does the effect of *sat* on *colgpa* differ by gender? Justify your answer.

C5 In Problem 2 in Chapter 4, we added the return on the firm's stock, *ros*, to a model explaining CEO salary; *ros* turned out to be insignificant. Now, define a dummy variable, *rosneg*, which is equal to one if $ros < 0$ and equal to zero if $ros \geq 0$. Use CEOSAL1 to estimate the model

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 rosneg + u.$$

Discuss the interpretation and statistical significance of $\hat{\beta}_3$.

C6 Use the data in SLEEP75 for this exercise. The equation of interest is

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u.$$

(i) Estimate this equation separately for men and women and report the results in the usual form. Are there notable differences in the two estimated equations?

(ii) Compute the Chow test for equality of the parameters in the sleep equation for men and women. Use the form of the test that adds *male* and the interaction terms *male*·*totwrk*, …, *male*·*yngkid* and uses the full set of observations. What are the relevant *df* for the test? Should you reject the null at the 5% level?

(iii) Now, allow for a different intercept for males and females and determine whether the interaction terms involving *male* are jointly significant.

(iv) Given the results from parts (ii) and (iii), what would be your final model?

C7 Use the data in WAGE1 for this exercise.

(i) Use equation (7.18) to estimate the gender differential when $educ = 12.5$. Compare this with the estimated differential when $educ = 0$.

(ii) Run the regression used to obtain (7.18), but with $female \cdot (educ - 12.5)$ replacing *female*·*educ*. How do you interpret the coefficient on *female* now?

(iii) Is the coefficient on *female* in part (ii) statistically significant? Compare this with (7.18) and comment.

C8 Use the data in LOANAPP for this exercise. The binary variable to be explained is *approve*, which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is *white*, a dummy variable equal to one if the applicant was white. The other applicants in the data set are black and Hispanic.

To test for discrimination in the mortgage loan market, a linear probability model can be used:

$$approve = \beta_0 + \beta_1 white + other\ factors.$$

(i) If there is discrimination against minorities, and the appropriate factors have been controlled for, what is the sign of β_1?

(ii) Regress *approve* on *white* and report the results in the usual form. Interpret the coefficient on *white*. Is it statistically significant? Is it practically large?

(iii) As controls, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, and *vr*. What happens to the coefficient on *white*? Is there still evidence of discrimination against nonwhites?

(iv) Now, allow the effect of race to interact with the variable measuring other obligations as a percentage of income (*obrat*). Is the interaction term significant?

(v) Using the model from part (iv), what is the effect of being white on the probability of approval when $obrat = 32$, which is roughly the mean value in the sample? Obtain a 95% confidence interval for this effect.

C9 There has been much interest in whether the presence of 401(k) pension plans, available to many U.S. workers, increases net savings. The data set 401KSUBS contains information on net financial assets (*nettfa*), family income (*inc*), a binary variable for eligibility in a 401(k) plan (*e401k*), and several other variables.

(i) What fraction of the families in the sample are eligible for participation in a 401(k) plan?

(ii) Estimate a linear probability model explaining 401(k) eligibility in terms of income, age, and gender. Include income and age in quadratic form, and report the results in the usual form.

(iii) Would you say that 401(k) eligibility is independent of income and age? What about gender? Explain.

(iv) Obtain the fitted values from the linear probability model estimated in part (ii). Are any fitted values negative or greater than one?

(v) Using the fitted values $\widehat{e401k}_i$ from part (iv), define $\widetilde{e401k}_i = 1$ if $\widehat{e401k}_i \geq .5$ and $\widetilde{e401k}_i = 0$ if $\widehat{e401k}_i < .5$. Out of 9,275 families, how many are predicted to be eligible for a 401(k) plan?

(vi) For the 5,638 families not eligible for a 401(k), what percentage of these are predicted not to have a 401(k), using the predictor $\widetilde{e401k}_i$? For the 3,637 families eligible for a 401(k) plan, what percentage are predicted to have one? (It is helpful if your econometrics package has a "tabulate" command.)

(vii) The overall percent correctly predicted is about 64.9%. Do you think this is a complete description of how well the model does, given your answers in part (vi)?

(viii) Add the variable *pira* as an explanatory variable to the linear probability model. Other things equal, if a family has someone with an individual retirement account, how much higher is the estimated probability that the family is eligible for a 401(k) plan? Is it statistically different from zero at the 10% level?

C10 Use the data in NBASAL for this exercise.

(i) Estimate a linear regression model relating points per game to experience in the league and position (guard, forward, or center). Include experience in quadratic form and use centers as the base group. Report the results in the usual form.

(ii) Why do you not include all three position dummy variables in part (i)?

(iii) Holding experience fixed, does a guard score more than a center? How much more? Is the difference statistically significant?

(iv) Now, add marital status to the equation. Holding position and experience fixed, are married players more productive (based on points per game)?

(v) Add interactions of marital status with both experience variables. In this expanded model, is there strong evidence that marital status affects points per game?

(vi) Estimate the model from part (iv) but use assists per game as the dependent variable. Are there any notable differences from part (iv)? Discuss.

C11 Use the data in 401KSUBS for this exercise.

(i) Compute the average, standard deviation, minimum, and maximum values of *nettfa* in the sample.

(ii) Test the hypothesis that average *nettfa* does not differ by 401(k) eligibility status; use a two-sided alternative. What is the dollar amount of the estimated difference?

(iii) From part (ii) of Computer Exercise C9, it is clear that *e401k* is not exogenous in a simple regression model; at a minimum, it changes by income and age. Estimate a multiple linear regression model for *nettfa* that includes income, age, and *e401k* as explanatory variables. The income and age variables should appear as quadratics. Now, what is the estimated dollar effect of 401(k) eligibility?

(iv) To the model estimated in part (iii), add the interactions $e401k \cdot (age - 41)$ and $e401k \cdot (age - 41)^2$. Note that the average age in the sample is about 41, so that in the new model, the coefficient on *e401k* is the estimated effect of 401(k) eligibility at the average age. Which interaction term is significant?

(v) Comparing the estimates from parts (iii) and (iv), do the estimated effects of 401(k) eligibility at age 41 differ much? Explain.

(vi) Now, drop the interaction terms from the model, but define five family size dummy variables: *fsize1, fsize2, fsize3, fsize4,* and *fsize5*. The variable *fsize5* is unity for families with five or more members. Include the family size dummies in the model estimated from part (iii); be sure to choose a base group. Are the family dummies significant at the 1% level?

(vii) Now, do a Chow test for the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 e401k + u$$

across the five family size categories, allowing for intercept differences. The restricted sum of squared residuals, SSR_r, is obtained from part (vi) because that regression assumes all slopes are the same. The unrestricted sum of squared residuals is $SSR_{ur} = SSR_1 + SSR_2 + \cdots + SSR_5$, where SSR_f is the sum of squared residuals for the equation estimated using only family size *f*. You should convince yourself that there are 30 parameters in the unrestricted model (5 intercepts plus 25 slopes) and 10 parameters in the restricted model (5 intercepts plus 5 slopes). Therefore, the number of restrictions being tested is $q = 20$, and the *df* for the unrestricted model is $9{,}275 - 30 = 9{,}245$.

C12 Use the data set in BEAUTY, which contains a subset of the variables (but more usable observations than in the regressions) reported by Hamermesh and Biddle (1994).

(i) Find the separate fractions of men and women that are classified as having above average looks. Are more people rated as having above average or below average looks?

(ii) Test the null hypothesis that the population fractions of above-average-looking women and men are the same. Report the one-sided *p*-value that the fraction is higher for women. (*Hint:* Estimating a simple linear probability model is easiest.)

(iii) Now estimate the model

$$\log(wage) = \beta_0 + \beta_1 belavg + \beta_2 abvavg + u$$

separately for men and women, and report the results in the usual form. In both cases, interpret the coefficient on *belavg*. Explain in words what the hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 < 0$ means, and find the *p*-values for men and women.

(iv) Is there convincing evidence that women with above average looks earn more than women with average looks? Explain.

(v) For both men and women, add the explanatory variables *educ*, *exper*, $exper^2$, *union*, *goodhlth*, *black*, *married*, *south*, *bigcity*, *smllcity*, and *service*. Do the effects of the "looks" variables change in important ways?

(vi) Use the SSR form of the Chow *F* statistic to test whether the slopes of the regression functions in part (v) differ across men and women. Be sure to allow for an intercept shift under the null.

C13 Use the data in APPLE to answer this question.

(i) Define a binary variable as $ecobuy = 1$ if $ecolbs > 0$ and $ecobuy = 0$ if $ecolbs = 0$. In other words, *ecobuy* indicates whether, at the prices given, a family would buy any ecologically friendly apples. What fraction of families claim they would buy ecolabeled apples?

(ii) Estimate the linear probability model

$$ecobuy = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc + \beta_4 hhsize + \beta_5 educ + \beta_6 age + u,$$

and report the results in the usual form. Carefully interpret the coefficients on the price variables.

(iii) Are the nonprice variables jointly significant in the LPM? (Use the usual F statistic, even though it is not valid when there is heteroskedasticity.) Which explanatory variable other than the price variables seems to have the most important effect on the decision to buy ecolabeled apples? Does this make sense to you?

(iv) In the model from part (ii), replace *faminc* with log(*faminc*). Which model fits the data better, using *faminc* or log(*faminc*)? Interpret the coefficient on log(*faminc*).

(v) In the estimation in part (iv), how many estimated probabilities are negative? How many are bigger than one? Should you be concerned?

(vi) For the estimation in part (iv), compute the percent correctly predicted for each outcome, $ecobuy = 0$ and $ecobuy = 1$. Which outcome is best predicted by the model?

C14 Use the data in CHARITY to answer this question. The variable *respond* is a dummy variable equal to one if a person responded with a contribution on the most recent mailing sent by a charitable organization. The variable *resplast* is a dummy variable equal to one if the person responded to the previous mailing, *avggift* is the average of past gifts (in Dutch guilders), and *propresp* is the proportion of times the person has responded to past mailings.

(i) Estimate a linear probability model relating *respond* to *resplast* and *avggift*. Report the results in the usual form, and interpret the coefficient on *resplast*.

(ii) Does the average value of past gifts seem to affect the probability of responding?

(iii) Add the variable *propresp* to the model, and interpret its coefficient. (Be careful here: an increase of one in *propresp* is the largest possible change.)

(iv) What happened to the coefficient on *resplast* when *propresp* was added to the regression? Does this make sense?

(v) Add *mailsyear*, the number of mailings per year, to the model. How big is its estimated effect? Why might this not be a good estimate of the causal effect of mailings on responding?

C15 Use the data in FERTIL2 to answer this question.

(i) Find the smallest and largest values of *children* in the sample. What is the average of *children*? Does any woman have exactly the average number of children?

(ii) What percentage of women have electricity in the home?

(iii) Compute the average of *children* for those without electricity and do the same for those with electricity. Comment on what you find. Test whether the population means are the same using a simple regression.

(iv) From part (iii), can you infer that having electricity "causes" women to have fewer children? Explain.

(v) Estimate a multiple regression model of the kind reported in equation (7.37), but add age^2, *urban*, and the three religious affiliation dummies. How does the estimated effect of having electricity compare with that in part (iii)? Is it still statistically significant?

(vi) To the equation in part (v), add an interaction between *electric* and *educ*. Is its coefficient statistically significant? What happens to the coefficient on *electric*?

(vii) The median and mode value for *educ* is 7. In the equation from part (vi), use the centered interaction term $electric \cdot (educ - 7)$ in place of $electric \cdot educ$. What happens to the coefficient on *electric* compared with part (vi)? Why? How does the coefficient on *electric* compare with that in part (v)?

C16 Use the data in CATHOLIC to answer this question.

(i) In the entire sample, what percentage of the students attend a Catholic high school? What is the average of *math*12 in the entire sample?

(ii) Run a simple regression of *math*12 on *cathhs* and report the results in the usual way. Interpret what you have found.

(iii) Now add the variables *lfaminc*, *motheduc*, and *fatheduc* to the regression from part (ii). How many observations are used in the regression? What happens to the coefficient on *cathhs*, along with its statistical significance?

(iv) Return to the simple regression of *math*12 on *cathhs*, but restrict the regression to observations used in the multiple regression from part (iii). Do any important conclusions change?

(v) To the multiple regression in part (iii), add interactions between *cathhs* and each of the other explanatory variables. Are the interaction terms individually or jointly significant?

(vi) What happens to the coefficient on *cathhs* in the regression from part (v). Explain why this coefficient is not very interesting.

(vii) Compute the average partial effect of *cathhs* in the model estimated in part (v). How does it compare with the coefficients on *cathhs* in parts (iii) and (v)?

C17 Use the data in JTRAIN98 to answer this question. The variable *unem*98 is a binary variable indicating whether a worker was unemployed in 1998. It can be used to measure the effectiveness of the job training program in reducing the probability of being unemployed.

(i) What percentage of workers was unemployed in 1998, after the job training program? How does this compare with the unemployment rate in 1996?

(ii) Run the simple regression *unem*98 on *train.* How do you interepret the coefficient on *train?* Is it statistically significant? Does it make sense to you?

(iii) Add to the regression in part (ii) the explanatory variables *earn*96, *educ, age*, and *married.* Now interpret the estimated training effect. Why does it differ so much from that in part (ii)?

(iv) Now perform full regression adjustment by running a regression with a full set of interactions, where all variables (except the training indicator) are centered around their sample means:

$$unem98_i \text{ on } train_i,\ earn96_i,\ educ_i,\ age_i,\ married,\ train_i \cdot (earn96_i - \overline{earn96}),$$
$$train_i \cdot (educ_i - \overline{educ}),\ train_i \cdot (age_i - \overline{age}),\ train_i \cdot (married_i - \overline{married}).$$

This regression uses all of the data. What ha ppens to the estimated average treatment effect of *train* compared with part (iii). Does its standard error change much?

(v) Are the interaction terms in part (iv) jointly significant?

(vi) Verify that you obtain exactly the same average treatment effect if you run two separate regressions and use the formula in equation (7.43). That is, run two separate regressions for the control and treated groups, obtain the fitted values $\widehat{unem98}_i^{(0)}$ and $\widehat{unem98}_i^{(1)}$ for everyone in the sample, and then compute

$$\hat{\tau}_{ura} = n^{-1}\sum_{i=1}^{n}[\widehat{unem98}_i^{(1)} - \widehat{unem98}_i^{(0)}].$$

Check this with the coefficient on *train* in part (iv). Which approach is more convenient for obtaining a standard error?

CHAPTER 8

Heteroskedasticity

The homoskedasticity assumption, introduced in Chapter 3 for multiple regression, states that the variance of the unobserved error, u, conditional on the explanatory variables, is constant. Homoskedasticity fails whenever the variance of the unobserved factors changes across different segments of the population, where the segments are determined by the different values of the explanatory variables. For example, in a savings equation, heteroskedasticity is present if the variance of the unobserved factors affecting savings increases with income.

In Chapters 4 and 5, we saw that homoskedasticity is needed to justify the usual t tests, F tests, and confidence intervals for OLS estimation of the linear regression model, even with large sample sizes. In this chapter, we discuss the available remedies when heteroskedasticity occurs, and we also show how to test for its presence. We begin by briefly reviewing the consequences of heteroskedasticity for ordinary least squares estimation.

8-1 Consequences of Heteroskedasticity for OLS

Consider again the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u. \qquad [8.1]$$

In Chapter 3, we proved unbiasedness of the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ under the first four Gauss-Markov assumptions, MLR.1 through MLR.4. In Chapter 5, we showed that the same four assumptions imply consistency of OLS. The homoskedasticity assumption MLR.5, stated in terms of the error variance as $\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$, played no role in showing whether OLS was unbiased or

consistent. It is important to remember that heteroskedasticity does not cause bias or inconsistency in the OLS estimators of the β_j, whereas something like omitting an important variable would have this effect.

The interpretation of our goodness-of-fit measures, R^2 and $\overline{R}^2$, is also unaffected by the presence of heteroskedasticity. Why? Recall from Section 6-3 that the usual R-squared and the adjusted R-squared are different ways of estimating the population R-squared, which is simply $1 - \sigma_u^2/\sigma_y^2$, where σ_u^2 is the population error variance and σ_y^2 is the population variance of y. The key point is that because both variances in the population R-squared are unconditional variances, the population R-squared is unaffected by the presence of heteroskedasticity in $\text{Var}(u|x_1, \ldots, x_k)$. Further, SSR/$n$ consistently estimates σ_u^2, and SST/n consistently estimates σ_y^2, whether or not $\text{Var}(u|x_1, \ldots, x_k)$ is constant. The same is true when we use the degrees of freedom adjustments. Therefore, R^2 and $\overline{R}^2$ are both consistent estimators of the population R-squared whether or not the homoskedasticity assumption holds.

If heteroskedasticity does not cause bias or inconsistency in the OLS estimators, why did we introduce it as one of the Gauss-Markov assumptions? Recall from Chapter 3 that the estimators of the *variances*, $\text{Var}(\hat{\beta}_j)$, are biased without the homoskedasticity assumption. Because the OLS standard errors are based directly on these variances, they are no longer valid for constructing confidence intervals and t statistics. The usual OLS t statistics do not have t distributions in the presence of heteroskedasticity, and the problem is not resolved by using large sample sizes. We will see this explicitly for the simple regression case in the next section, where we derive the variance of the OLS slope estimator under heteroskedasticity and propose a valid estimator in the presence of heteroskedasticity. Similarly, F statistics are no longer F distributed, and the LM statistic no longer has an asymptotic chi-square distribution. In summary, the statistics we used to test hypotheses under the Gauss-Markov assumptions are not valid in the presence of heteroskedasticity.

We also know that the Gauss-Markov Theorem, which says that OLS is best linear unbiased, relies crucially on the homoskedasticity assumption. If $\text{Var}(u|\mathbf{x})$ is not constant, OLS is no longer BLUE. In addition, OLS is no longer asymptotically efficient in the class of estimators described in Theorem 5.3. As we will see in Section 8-4, it is possible to find estimators that are more efficient than OLS in the presence of heteroskedasticity (although it requires knowing the form of the heteroskedasticity). With relatively large sample sizes, it might not be so important to obtain an efficient estimator. In the next section, we show how the usual OLS test statistics can be modified so that they are valid, at least asymptotically.

8-2 Heteroskedasticity-Robust Inference after OLS Estimation

Because testing hypotheses is such an important component of any econometric analysis and the usual OLS inference is generally faulty in the presence of heteroskedasticity, we must decide if we should entirely abandon OLS. Fortunately, OLS is still useful. In the last two decades, econometricians have learned how to adjust standard errors and t, F, and LM statistics so that they are valid in the presence of **heteroskedasticity of unknown form**. This is very convenient because it means we can report new statistics that work regardless of the kind of heteroskedasticity present in the population. The methods in this section are known as *heteroskedasticity-robust* procedures because they are valid—at least in large samples—whether or not the errors have constant variance, and we do not need to know which is the case.

We begin by sketching how the variances, $\text{Var}(\hat{\beta}_j)$, can be estimated in the presence of heteroskedasticity. A careful derivation of the theory is well beyond the scope of this text, but the application of heteroskedasticity-robust methods is very easy now because many statistics and econometrics packages compute these statistics as an option.

First, consider the model with a single independent variable, where we include an i subscript for emphasis:

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

We assume throughout that the first four Gauss-Markov assumptions hold. If the errors contain heteroskedasticity, then

$$\text{Var}(u_i|x_i) = \sigma_i^2,$$

where we put an i subscript on σ^2 to indicate that the variance of the error depends upon the particular value of x_i.

Write the OLS estimator as

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Under Assumptions MLR.1 through MLR.4 (that is, without the homoskedasticity assumption), and conditioning on the values x_i in the sample, we can use the same arguments from Chapter 2 to show that

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{\text{SST}_x^2}, \qquad [8.2]$$

where $\text{SST}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ is the total sum of squares of the x_i. When $\sigma_i^2 = \sigma^2$ for all i, this formula reduces to the usual form, σ^2/SST_x. Equation (8.2) explicitly shows that, for the simple regression case, the variance formula derived under homoskedasticity is no longer valid when heteroskedasticity is present.

Because the standard error of $\hat{\beta}_1$ is based directly on estimating $\text{Var}(\hat{\beta}_1)$, we need a way to estimate equation (8.2) when heteroskedasticity is present. White (1980) showed how this can be done. Let $\hat{u}_i$ denote the OLS residuals from the initial regression of y on x. Then, a valid estimator of $\text{Var}(\hat{\beta}_1)$, for heteroskedasticity of *any* form (including homoskedasticity), is

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\hat{u}_i^2}{\text{SST}_x^2}, \qquad [8.3]$$

which is easily computed from the data after the OLS regression.

In what sense is (8.3) a valid estimator of $\text{Var}(\hat{\beta}_1)$? This is pretty subtle. Briefly, it can be shown that when equation (8.3) is multiplied by the sample size n, it converges in probability to $\text{E}[(x_i - \mu_x)^2u_i^2]/(\sigma_x^2)^2$, which is the probability limit of n times (8.2). Ultimately, this is what is necessary for justifying the use of standard errors to construct confidence intervals and t statistics. The law of large numbers and the central limit theorem play key roles in establishing these convergences. You can refer to White's original paper for details, but that paper is quite technical. See also Wooldridge (2010, Chapter 4).

A similar formula works in the general multiple regression model

$$y = \beta_0 + \beta_1x_1 + \cdots + \beta_kx_k + u.$$

It can be shown that a valid estimator of $\text{Var}(\hat{\beta}_j)$, under Assumptions MLR.1 through MLR.4, is

$$\widehat{\text{Var}}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n}\hat{r}_{ij}^2\hat{u}_i^2}{\text{SSR}_j^2}, \qquad [8.4]$$

where $\hat{r}_{ij}$ denotes the i^{th} residual from regressing x_j on all other independent variables, and SSR_j is the sum of squared residuals from this regression (see Section 3-2 for the partialling out representation of the OLS estimates). The square root of the quantity in (8.4) is called the **heteroskedasticity-robust standard error** for $\hat{\beta}_j$. In econometrics, these robust standard errors are usually attributed to White (1980). Earlier works in statistics, notably those by Eicker (1967) and Huber (1967), pointed to the possibility of obtaining such robust standard errors. In applied work, these are sometimes called *White*, *Huber*, or *Eicker standard errors* (or some hyphenated combination of these names). We will just refer to them as *heteroskedasticity-robust standard errors*, or even just *robust standard errors* when the context is clear.

Sometimes, as a degrees of freedom correction, (8.4) is multiplied by $n/(n - k - 1)$ before taking the square root. The reasoning for this adjustment is that, if the squared OLS residuals $\hat{u}_i^2$ were the same for all observations i—the strongest possible form of homoskedasticity in a sample—we would get the usual OLS standard errors. Other modifications of (8.4) are studied in MacKinnon and White (1985). Because all forms have only asymptotic justification and they are asymptotically equivalent, no form is uniformly preferred above all others. Typically, we use whatever form is computed by the regression package at hand.

Once heteroskedasticity-robust standard errors are obtained, it is simple to construct a **heteroskedasticity-robust *t* statistic**. Recall that the general form of the t statistic is

$$t = \frac{\textit{estimate} - \textit{hypothesized value}}{\textit{standard error}}. \qquad [8.5]$$

Because we are still using the OLS estimates and we have chosen the hypothesized value ahead of time, the only difference between the usual OLS t statistic and the heteroskedasticity-robust t statistic is in how the standard error in the denominator is computed.

The term SSR_j in equation (8.4) can be replaced with $\text{SST}_j(1 - R_j^2)$, where SST_j is the total sum of squares of x_j and R_j^2 is the usual R-squared from regressing x_j on all other explanatory variables. [We implicitly used this equivalence in deriving equation (3.51).] Consequently, little sample variation in x_j, or a strong linear relationship between x_j and the other explanatory variables—that is, multicollinearity—can cause the heteroskedasticity-robust standard errors to be large. We discussed these issues with the usual OLS standard errors in Section 3-4.

EXAMPLE 8.1 **Log Wage Equation with Heteroskedasticity-Robust Standard Errors**

We estimate the model in Example 7.6, but we report the heteroskedasticity-robust standard errors along with the usual OLS standard errors. Some of the estimates are reported to more digits so that we can compare the usual standard errors with the heteroskedasticity-robust standard errors:

$$\begin{aligned}\widehat{\log(wage)} = {} & .321 + .213\,marrmale - .198\,marrfem - .110\,singfem \\ & \quad (.100)\ (.055) \qquad (.058) \qquad (.056) \\ & \quad [.109]\ [.057] \qquad [.058] \qquad [.057] \\ & + .0789\,educ + .0268\,exper - .00054\,exper^2 \\ & \quad (.0067) \qquad (.0052) \qquad (.00011) \\ & \quad [.0074] \qquad [.0051] \qquad [.00011] \\ & + .0291\,tenure - .00053\,tenure^2 \\ & \quad (.0068) \qquad (.00023) \\ & \quad [.0069] \qquad [.00024] \\ n = 526,\ & R^2 = .461. \end{aligned} \qquad [8.6]$$

The usual OLS standard errors are in parentheses, ( ), below the corresponding OLS estimate, and the heteroskedasticity-robust standard errors are in brackets, [ ]. The numbers in brackets are the only new things, as the equation is still estimated by OLS.

Several things are apparent from equation (8.6). First, in this particular application, any variable that was statistically significant using the usual t statistic is still statistically significant using the heteroskedasticity-robust t statistic. This occurs because the two sets of standard errors are not very different. (The associated p-values will differ slightly because the robust t statistics are not identical to the usual, nonrobust t statistics.) The largest relative change in standard errors is for the coefficient on *educ:* the usual standard error is .0067, and the robust standard error is .0074. Still, the robust standard error implies a robust t statistic above 10.

Equation (8.6) also shows that the robust standard errors can be either larger or smaller than the usual standard errors. For example, the robust standard error on *exper* is .0051, whereas the usual standard error is .0055. We do not know which will be larger ahead of time. As an empirical matter, the robust standard errors are often found to be larger than the usual standard errors.

Before leaving this example, we must emphasize that we do not know, at this point, whether heteroskedasticity is even present in the population model underlying equation (8.6). All we have done is report, along with the usual standard errors, those that are valid (asymptotically) whether or not heteroskedasticity is present. We can see that no important conclusions are overturned by using the robust standard errors in this example. This often happens in applied work, but in other cases, the differences between the usual and robust standard errors are much larger. As an example of where the differences are substantial, see Computer Exercise C2.

At this point, you may be asking the following question: if the heteroskedasticity-robust standard errors are valid more often than the usual OLS standard errors, why do we bother with the usual standard errors at all? This is a sensible question. One reason the usual standard errors are still used in cross-sectional work is that, if the homoskedasticity assumption holds and the errors are normally distributed, then the usual t statistics have *exact* t distributions, regardless of the sample size (see Chapter 4). The robust standard errors and robust t statistics are justified only as the sample size becomes large, even if the CLM assumptions are true. With small sample sizes, the robust t statistics can have distributions that are not very close to the t distribution, and that could throw off our inference.

In large sample sizes, we can make a case for always reporting only the heteroskedasticity-robust standard errors in cross-sectional applications, and this practice is being followed more and more in applied work. It is also common to report both standard errors, as in equation (8.6), so that a reader can determine whether any conclusions are sensitive to the standard error in use.

It is also possible to obtain F and LM statistics that are robust to heteroskedasticity of an unknown, arbitrary form. The **heteroskedasticity-robust *F* statistic** (or a simple transformation of it) is also called a *heteroskedasticity-robust Wald statistic*. A general treatment of the Wald statistic requires matrix algebra and is sketched in Advanced Treatment E; see Wooldridge (2010, Chapter 4) for a more detailed treatment. Nevertheless, using heteroskedasticity-robust statistics for multiple exclusion restrictions is straightforward because many econometrics packages now compute such statistics routinely.

EXAMPLE 8.2 Heteroskedasticity-Robust *F* Statistic

Using the data for the spring semester in GPA3, we estimate the following equation:

$$\begin{aligned}\widehat{cumgpa} = {} & 1.47 + .00114\,sat - .00857\,hsperc + .00250\,tothrs \\ & \quad (.23)\ (.00018)\quad (.00124)\quad\ (.00073) \\ & \quad [.22]\ [.00019]\quad [.00140]\quad\ [.00073] \\ & + .303\,female - .128\,black - .059\,white \\ & \quad (.059)\quad\ (.147)\quad\ (.141) \\ & \quad [.059]\quad\ [.118]\quad\ [.110] \\ n = 366,\ & R^2 = .4006,\ \overline{R}^2 = .3905.\end{aligned} \tag{8.7}$$

Again, the differences between the usual standard errors and the heteroskedasticity-robust standard errors are not very big, and use of the robust t statistics does not change the statistical significance of any independent variable. Joint significance tests are not much affected either. Suppose we wish to test the null hypothesis that, after the other factors are controlled for, there are no differences in *cumgpa* by race. This is stated as H_0: $\beta_{black} = 0, \beta_{white} = 0$. The usual F statistic is easily obtained, once we have the R-squared from the restricted model; this turns out to be .3983. The F statistic is then $[(.4006 - .3983)/(1 - .4006)](359/2) \approx .69$. If heteroskedasticity is present, this version of the test is invalid. The heteroskedasticity-robust version has no simple form, but it can be computed using certain statistical packages. The value of the heteroskedasticity-robust F statistic turns out to be .75, which differs only slightly from the nonrobust version. The p-value for the robust test is .474, which is not close to standard significance levels. We fail to reject the null hypothesis using either test.

Because the usual sum of squared residuals form of the F statistic is not valid under heteroskedasticity, we must be careful in computing a Chow test of common coefficients across two groups. The form of the statistic in equation (7.24) is not valid if heteroskedasticity is present, including the simple case where the error variance differs across the two groups. Instead, we can obtain a heteroskedasticity-robust Chow test by including a dummy variable distinguishing the two groups along with interactions between that dummy variable and all other explanatory variables. We can then test whether there is no difference in the two regression functions—by testing that the coefficients on the dummy variable and all interactions are zero—or just test whether the slopes are all the same, in which case we leave the coefficient on the dummy variable unrestricted. See Computer Exercise C14 for an example.

8-2a Computing Heteroskedasticity-Robust *LM* Tests

GOING FURTHER 8.1

Evaluate the following statement: The heteroskedasticity-robust standard errors are always bigger than the usual standard errors.

Not all regression packages compute F statistics that are robust to heteroskedasticity. Therefore, it is sometimes convenient to have a way of obtaining a test of multiple exclusion restrictions that is robust to heteroskedasticity and does not require a particular kind of econometric software. It turns out that a **heteroskedasticity-robust *LM* statistic** is easily obtained using virtually any regression package.

To illustrate computation of the robust *LM* statistic, consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u,$$

and suppose we would like to test H_0: $\beta_4 = 0, \ \beta_5 = 0$. To obtain the usual *LM* statistic, we would first estimate the restricted model (that is, the model without x_4 and x_5) to obtain the residuals, $\tilde{u}$. Then, we would regress $\tilde{u}$ on all of the independent variables and the $LM = n \cdot R_{\tilde{u}}^2$, where $R_{\tilde{u}}^2$ is the usual R-squared from this regression.

Obtaining a version that is robust to heteroskedasticity requires more work. One way to compute the statistic requires only OLS regressions. We need the residuals, say, $\tilde{r}_1$, from the regression of x_4 on x_1, x_2, x_3. Also, we need the residuals, say, $\tilde{r}_2$, from the regression of x_5 on x_1, x_2, x_3. Thus, we regress each of the independent variables excluded under the null on all of the included independent variables. We keep the residuals each time. The final step appears odd, but it is, after all, just a computational device. Run the regression of

$$1 \text{ on } \tilde{r}_1\tilde{u}, \tilde{r}_2\tilde{u}, \qquad [8.8]$$

without an intercept. Yes, we actually define a dependent variable equal to the value one for all observations. We regress this onto the products $\tilde{r}_1\tilde{u}$ and $\tilde{r}_2\tilde{u}$. The robust *LM* statistic turns out to be $n - SSR_1$, where SSR_1 is just the usual sum of squared residuals from regression (8.8).

The reason this works is somewhat technical. Basically, this is doing for the *LM* test what the robust standard errors do for the *t* test. [See Wooldridge (1991b) or Davidson and MacKinnon (1993) for a more detailed discussion.]

We now summarize the computation of the heteroskedasticity-robust *LM* statistic in the general case.

A Heteroskedasticity-Robust *LM* Statistic:

1. Obtain the residuals $\tilde{u}$ from the restricted model.
2. Regress each of the independent variables excluded under the null on all of the included independent variables; if there are q excluded variables, this leads to q sets of residuals $(\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_q)$.
3. Find the products of each $\tilde{r}_j$ and $\tilde{u}$ (for all observations).
4. Run the regression of 1 on $\tilde{r}_1\tilde{u}, \tilde{r}_2\tilde{u}, \ldots, \tilde{r}_q\tilde{u}$, without an intercept. The heteroskedasticity-robust *LM* statistic is $n - \text{SSR}_1$, where SSR_1 is just the usual sum of squared residuals from this final regression. Under H_0, *LM* is distributed approximately as χ^2_q.

Once the robust *LM* statistic is obtained, the rejection rule and computation of *p*-values are the same as for the usual *LM* statistic in Section 5-2.

EXAMPLE 8.3 Heteroskedasticity-Robust *LM* Statistic

We use the data in CRIME1 to test whether the average sentence length served for past convictions affects the number of arrests in the current year (1986). The estimated model is

$$\begin{aligned}\widehat{narr86} = {} & .561 - .136\,pcnv + .0178\,avgsen - .00052\,avgsen^2 \\ & \quad (.036)\ (.040)\quad (.0097)\quad (.00030) \\ & \quad [.040]\ [.034]\quad [.0101]\quad [.00021] \\ & - .0394\,ptime86 - .0505\,qemp86 - .00148\,inc86 \\ & \quad (.0087)\quad (.0144)\quad (.00034) \\ & \quad [.0062]\quad [.0142]\quad [.00023] \\ & + .325\,black + .193\,hispan \\ & \quad (.045)\quad (.040) \\ & \quad [.058]\quad [.040] \\ n = {} & 2{,}725,\ R^2 = .0728.\end{aligned} \qquad [8.9]$$

In this example, there are more substantial differences between some of the usual standard errors and the robust standard errors. For example, the usual *t* statistic on $avgsen^2$ is about −1.73, while the robust *t* statistic is about −2.48. Thus, $avgsen^2$ is more significant using the robust standard error.

The effect of *avgsen* on *narr86* is somewhat difficult to reconcile. Because the relationship is quadratic, we can figure out where *avgsen* has a positive effect on *narr86* and where the effect becomes negative. The turning point is .0178/[2(.00052)] ≈ 17.12; recall that this is measured in months. Literally, this means that *narr86* is positively related to *avgsen* when *avgsen* is less than 17 months; then *avgsen* has the expected deterrent effect after 17 months.

To see whether average sentence length has a statistically significant effect on *narr86*, we must test the joint hypothesis H_0: $\beta_{avgsen} = 0, \beta_{avgsen^2} = 0$. Using the usual *LM* statistic (see Section 5-2), we obtain $LM = 3.54$; in a chi-square distribution with two *df*, this yields a *p*-value = .170. Thus, we do not reject H_0 at even the 15% level. The heteroskedasticity-robust *LM* statistic is $LM = 4.00$ (rounded to two decimal places), with a *p*-value = .135. This is still not very strong evidence against H_0; *avgsen* does not appear to have a strong effect on *narr86*. [Incidentally, when *avgsen* appears alone in (8.9), that is, without the quadratic term, its usual *t* statistic is .658, and its robust *t* statistic is .592.]

8-3 Testing for Heteroskedasticity

The heteroskedasticity-robust standard errors provide a simple method for computing t statistics that are asymptotically t distributed whether or not heteroskedasticity is present. We have also seen that heteroskedasticity-robust F and LM statistics are available. Implementing these tests does not require knowing whether or not heteroskedasticity is present. Nevertheless, there are still some good reasons for having simple tests that can detect its presence. First, as we mentioned in the previous section, the usual t statistics have exact t distributions under the classical linear model assumptions. For this reason, many economists still prefer to see the usual OLS standard errors and test statistics reported, unless there is evidence of heteroskedasticity. Second, if heteroskedasticity is present, the OLS estimator is no longer the best linear unbiased estimator. As we will see in Section 8-4, it is possible to obtain a better estimator than OLS when the form of heteroskedasticity is known.

Many tests for heteroskedasticity have been suggested over the years. Some of them, while having the ability to detect heteroskedasticity, do not directly test the assumption that the variance of the error does not depend upon the independent variables. We will restrict ourselves to more modern tests, which detect the kind of heteroskedasticity that invalidates the usual OLS statistics. This also has the benefit of putting all tests in the same framework.

As usual, we start with the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u, \qquad [8.10]$$

where Assumptions MLR.1 through MLR.4 are maintained in this section. In particular, we assume that $\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$, so that OLS is unbiased and consistent.

We take the null hypothesis to be that Assumption MLR.5 is true:

$$\mathrm{H}_0\colon \mathrm{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2. \qquad [8.11]$$

That is, we assume that the ideal assumption of homoskedasticity holds, and we require the data to tell us otherwise. If we cannot reject (8.11) at a sufficiently small significance level, we usually conclude that heteroskedasticity is not a problem. However, remember that we never accept H_0; we simply fail to reject it.

Because we are assuming that u has a zero conditional expectation, $\mathrm{Var}(u|\mathbf{x}) = \mathrm{E}(u^2|\mathbf{x})$, and so the null hypothesis of homoskedasticity is equivalent to

$$\mathrm{H}_0\colon \mathrm{E}(u^2|x_1, x_2, \ldots, x_k) = \mathrm{E}(u^2) = \sigma^2.$$

This shows that, in order to test for violation of the homoskedasticity assumption, we want to test whether u^2 is related (in expected value) to one or more of the explanatory variables. If H_0 is false, the expected value of u^2, given the independent variables, can be virtually any function of the x_j. A simple approach is to assume a linear function:

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + v, \qquad [8.12]$$

where v is an error term with mean zero given the x_j. Pay close attention to the dependent variable in this equation: it is the *square* of the error in the original regression equation, (8.10). The null hypothesis of homoskedasticity is

$$\mathrm{H}_0\colon \delta_1 = \delta_2 = \cdots = \delta_k = 0. \qquad [8.13]$$

Under the null hypothesis, it is often reasonable to assume that the error in (8.12), v, is independent of $x_1, x_2, \ldots, x_k$. Then, we know from Section 5-2 that either the F or LM statistics for the overall significance of the independent variables in explaining u^2 can be used to test (8.13). Both statistics would have asymptotic justification, even though u^2 cannot be normally distributed. (For example, if u is normally distributed, then u^2/σ^2 is distributed as χ^2_1.) If we could observe the u^2 in the sample, then we could easily compute this statistic by running the OLS regression of u^2 on $x_1, x_2, \ldots, x_k$, using all n observations.

As we have emphasized before, we never know the actual errors in the population model, but we do have estimates of them: the OLS residual, $\hat{u}_i$, is an estimate of the error u_i for observation i. Thus, we can estimate the equation

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + error \qquad [8.14]$$

and compute the F or LM statistics for the joint significance of $x_1, \ldots, x_k$. It turns out that using the OLS residuals in place of the errors does not affect the large sample distribution of the F or LM statistics, although showing this is pretty complicated.

The F and LM statistics both depend on the R-squared from regression (8.14); call this $R^2_{\hat{u}^2}$ to distinguish it from the R-squared in estimating equation (8.10). Then, the F statistic is

$$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n - k - 1)}, \qquad [8.15]$$

where k is the number of regressors in (8.14); this is the same number of independent variables in (8.10). Computing (8.15) by hand is rarely necessary, because most regression packages automatically compute the F statistic for overall significance of a regression. This F statistic has (approximately) an $F_{k,\,n-k-1}$ distribution under the null hypothesis of homoskedasticity.

The LM statistic for heteroskedasticity is just the sample size times the R-squared from (8.14):

$$LM = n \cdot R^2_{\hat{u}^2}. \qquad [8.16]$$

Under the null hypothesis, LM is distributed asymptotically as χ^2_k. This is also very easy to obtain after running regression (8.14).

The LM version of the test is typically called the **Breusch-Pagan test for heteroskedasticity (BP test)**. Breusch and Pagan (1979) suggested a different form of the test that assumes the errors are normally distributed. Koenker (1981) suggested the form of the LM statistic in (8.16), and it is generally preferred due to its greater applicability.

We summarize the steps for testing for heteroskedasticity using the BP test:

The Breusch-Pagan Test for Heteroskedasticity:

1. Estimate the model (8.10) by OLS, as usual. Obtain the squared OLS residuals, $\hat{u}^2$ (one for each observation).
2. Run the regression in (8.14). Keep the R-squared from this regression, $R^2_{\hat{u}^2}$.
3. Form either the F statistic or the LM statistic and compute the p-value (using the $F_{k,\,n-k-1}$ distribution in the former case and the χ^2_k distribution in the latter case). If the p-value is sufficiently small, that is, below the chosen significance level, then we reject the null hypothesis of homoskedasticity.

If the BP test results in a small enough p-value, some corrective measure should be taken. One possibility is to just use the heteroskedasticity-robust standard errors and test statistics discussed in the previous section. Another possibility is discussed in Section 8-4.

EXAMPLE 8.4 Heteroskedasticity in Housing Price Equations

We use the data in HPRICE1 to test for heteroskedasticity in a simple housing price equation. The estimated equation using the levels of all variables is

$$\widehat{price} = -21.77 + .00207\, lotsize + .123\, sqrft + 13.85\, bdrms \qquad [8.17]$$
$$\qquad\quad (29.48)\ (.00064)\qquad (.013)\qquad (9.01)$$
$$n = 88,\ R^2 = .672.$$

This equation tells us *nothing* about whether the error in the population model is heteroskedastic. We need to regress the squared OLS residuals on the independent variables. The *R*-squared from the regression of $\hat{u}^2$ on *lotsize*, *sqrft*, and *bdrms* is $R^2_{\hat{u}^2} = .1601$. With $n = 88$ and $k = 3$, this produces an F statistic for significance of the independent variables of $F = [.1601/(1 - .1601)](84/3) \approx 5.34$. The associated p-value is .002, which is strong evidence against the null. The LM statistic is $88(.1601) \approx 14.09$; this gives a p-*value* $\approx .0028$ (using the χ^2_3 distribution), giving essentially the same conclusion as the F statistic. This means that the usual standard errors reported in (8.17) are not reliable.

In Chapter 6, we mentioned that one benefit of using the logarithmic functional form for the dependent variable is that heteroskedasticity is often reduced. In the current application, let us put *price*, *lotsize*, and *sqrft* in logarithmic form, so that the elasticities of *price*, with respect to *lotsize* and *sqrft*, are constant. The estimated equation is

$$\widehat{\log(price)} = -1.30 + .168\log(lotsize) + .700\log(sqrft) + 0.37\,bdrms \quad [8.18]$$
$$(.65)\ (.038)\qquad (.093)\qquad (.028)$$
$$n = 88, R^2 = .643.$$

Regressing the squared OLS residuals from this regression on log(*lotsize*), log(*sqrft*), and *bdrms* gives $R^2_{\hat{u}^2} = .0480$. Thus, $F = 1.41$ (p-value $= .245$), and $LM = 4.22$ (p-value $= .239$). Therefore, we fail to reject the null hypothesis of homoskedasticity in the model with the logarithmic functional forms. The occurrence of less heteroskedasticity with the dependent variable in logarithmic form has been noticed in many empirical applications.

GOING FURTHER 8.2

Consider wage equation (7.11), where you think that the conditional variance of log(*wage*) does not depend on *educ*, *exper*, or *tenure*. However, you are worried that the variance of log(*wage*) differs across the four demographic groups of married males, married females, single males, and single females. What regression would you run to test for heteroskedasticity? What are the degrees of freedom in the *F* test?

If we suspect that heteroskedasticity depends only upon certain independent variables, we can easily modify the Breusch-Pagan test: we simply regress $\hat{u}^2$ on whatever independent variables we choose and carry out the appropriate F or LM test. Remember that the appropriate degrees of freedom depends upon the number of independent variables in the regression with $\hat{u}^2$ as the dependent variable; the number of independent variables showing up in equation (8.10) is irrelevant.

If the squared residuals are regressed on only a single independent variable, the test for heteroskedasticity is just the usual t statistic on the variable. A significant t statistic suggests that heteroskedasticity is a problem.

8-3a The White Test for Heteroskedasticity

In Chapter 5, we showed that the usual OLS standard errors and test statistics are asymptotically valid, provided all of the Gauss-Markov assumptions hold. It turns out that the homoskedasticity assumption, $\text{Var}(u_1|x_1, \ldots, x_k) = \sigma^2$, can be replaced with the weaker assumption that the squared error, u^2, is *uncorrelated* with all the independent variables (x_j), the squares of the independent variables (x_j^2), and all the cross products ($x_j x_h$ for $j \neq h$). This observation motivated White (1980) to propose a test for heteroskedasticity that adds the squares and cross products of all the independent variables to equation (8.14). The test is explicitly intended to test for forms of heteroskedasticity that invalidate the usual OLS standard errors and test statistics.

When the model contains $k = 3$ independent variables, the White test is based on an estimation of

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error. \quad [8.19]$$

Compared with the Breusch-Pagan test, this equation has six more regressors. The **White test for heteroskedasticity** is the *LM* statistic for testing that all of the δ_j in equation (8.19) are zero, except for the intercept. Thus, nine restrictions are being tested in this case. We can also use an *F* test of this hypothesis; both tests have asymptotic justification.

With only three independent variables in the original model, equation (8.19) has nine independent variables. With six independent variables in the original model, the White regression would generally involve 27 regressors (unless some are redundant). This abundance of regressors is a weakness in the pure form of the White test: it uses many degrees of freedom for models with just a moderate number of independent variables.

It is possible to obtain a test that is easier to implement than the White test and more conserving on degrees of freedom. To create the test, recall that the difference between the White and Breusch-Pagan tests is that the former includes the squares and cross products of the independent variables. We can preserve the spirit of the White test while conserving on degrees of freedom by using the OLS fitted values in a test for heteroskedasticity. Remember that the fitted values are defined, for each observation *i*, by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}.$$

These are just linear functions of the independent variables. If we square the fitted values, we get a particular function of all the squares and cross products of the independent variables. This suggests testing for heteroskedasticity by estimating the equation

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error, \quad [8.20]$$

where $\hat{y}$ stands for the fitted values. It is important not to confuse $\hat{y}$ and y in this equation. We use the fitted values because they are functions of the independent variables (and the estimated parameters); using y in (8.20) does not produce a valid test for heteroskedasticity.

We can use the *F* or *LM* statistic for the null hypothesis H_0: $\delta_1 = 0, \delta_2 = 0$ in equation (8.20). This results in two restrictions in testing the null of homoskedasticity, regardless of the number of independent variables in the original model. Conserving on degrees of freedom in this way is often a good idea, and it also makes the test easy to implement.

Because $\hat{y}$ is an estimate of the expected value of y, given the x_j, using (8.20) to test for heteroskedasticity is useful in cases where the variance is thought to change with the level of the expected value, $E(y|\mathbf{x})$. The test from (8.20) can be viewed as a special case of the White test, as equation (8.20) can be shown to impose restrictions on the parameters in equation (8.19).

A Special Case of the White Test for Heteroskedasticity:

1. Estimate the model (8.10) by OLS, as usual. Obtain the OLS residuals $\hat{u}$ and the fitted values $\hat{y}$. Compute the squared OLS residuals $\hat{u}^2$ and the squared fitted values $\hat{y}^2$.
2. Run the regression in equation (8.20). Keep the *R*-squared from this regression, $R^2_{\hat{u}^2}$.
3. Form either the *F* or *LM* statistic and compute the *p*-value (using the $F_{2,n-3}$ distribution in the former case and the χ^2_2 distribution in the latter case).

EXAMPLE 8.5 **Special Form of the White Test in the Log Housing Price Equation**

We apply the special case of the White test to equation (8.18), where we use the *LM* form of the statistic. The important thing to remember is that the chi-square distribution always has two *df*. The regression of $\hat{u}^2$ on $\widehat{lprice}$, $(\widehat{lprice})^2$, where $\widehat{lprice}$ denotes the fitted values from (8.18), produces $R^2_{\hat{u}^2} = .0392$; thus, $LM = 88(.0392) \approx 3.45$, and the p-value $= .178$. This is stronger evidence of heteroskedasticity than is provided by the Breusch-Pagan test, but we still fail to reject homoskedasticity at even the 15% level.

Before leaving this section, we should discuss one important caveat. We have interpreted a rejection using one of the heteroskedasticity tests as evidence of heteroskedasticity. This is appropriate provided we maintain Assumptions MLR.1 through MLR.4. But, if MLR.4 is violated—in particular, if the functional form of $\mathrm{E}(y|\mathbf{x})$ is misspecified—then a test for heteroskedasticity can reject H_0, even if $\mathrm{Var}(y|\mathbf{x})$ is constant. For example, if we omit one or more quadratic terms in a regression model or use the level model when we should use the log, a test for heteroskedasticity can be significant. This has led some economists to view tests for heteroskedasticity as general misspecification tests. However, there are better, more direct tests for functional form misspecification, and we will cover some of them in Section 9-1. It is better to use explicit tests for functional form first, as functional form misspecification is more important than heteroskedasticity. Then, once we are satisfied with the functional form, we can test for heteroskedasticity.

8-4 Weighted Least Squares Estimation

If heteroskedasticity is detected using one of the tests in Section 8-3, we know from Section 8-2 that one possible response is to use heteroskedasticity-robust statistics after estimation by OLS. Before the development of heteroskedasticity-robust statistics, the response to a finding of heteroskedasticity was to specify its form and use a *weighted least squares* method, which we develop in this section. As we will argue, if we have correctly specified the form of the variance (as a function of explanatory variables), then weighted least squares (WLS) is more efficient than OLS, and WLS leads to new t and F statistics that have t and F distributions. We will also discuss the implications of using the wrong form of the variance in the WLS procedure.

8-4a The Heteroskedasticity Is Known up to a Multiplicative Constant

Let $\mathbf{x}$ denote all the explanatory variables in equation (8.10) and assume that

$$\mathrm{Var}(u|\mathbf{x}) = \sigma^2 h(\mathbf{x}), \tag{8.21}$$

where $h(\mathbf{x})$ is some function of the explanatory variables that determines the heteroskedasticity. Because variances must be positive, $h(\mathbf{x}) > 0$ for all possible values of the independent variables. For now, we assume that the function $h(\mathbf{x})$ is known. The population parameter σ^2 is unknown, but we will be able to estimate it from a data sample.

For a random drawing from the population, we can write $\sigma_i^2 = \mathrm{Var}(u_i|\mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i) = \sigma^2 h_i$, where we again use the notation $\mathbf{x}_i$ to denote all independent variables for observation i, and h_i changes with each observation because the independent variables change across observations. For example, consider the simple savings function

$$sav_i = \beta_0 + \beta_i inc_i + u_i \tag{8.22}$$

$$\mathrm{Var}(u_i|inc_i) = \sigma^2 inc_i. \tag{8.23}$$

Here, $h(x) = h(inc) = inc$: the variance of the error is proportional to the level of income. This means that, as income increases, the variability in savings increases. (If $\beta_1 > 0$, the expected value of savings also increases with income.) Because *inc* is always positive, the variance in equation (8.23) is always guaranteed to be positive. The standard deviation of u_i, conditional on inc_i, is $\sigma\sqrt{inc_i}$.

How can we use the information in equation (8.21) to estimate the β_j? Essentially, we take the original equation,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i, \qquad [8.24]$$

which contains heteroskedastic errors, and transform it into an equation that has homoskedastic errors (and satisfies the other Gauss-Markov assumptions). Because h_i is just a function of $\mathbf{x}_i$, $u_i/\sqrt{h_i}$ has a zero expected value conditional on $\mathbf{x}_i$. Further, because $\text{Var}(u_i|\mathbf{x}_i) = \text{E}(u_i^2|\mathbf{x}_i) = \sigma^2 h_i$, the variance of $u_i/\sqrt{h_i}$ (conditional on $\mathbf{x}_i$) is σ^2:

$$\text{E}[(u_i/\sqrt{h_i})^2] = \text{E}(u_i^2)/h_i = (\sigma^2 h_i)/h_i = \sigma^2,$$

where we have suppressed the conditioning on $\mathbf{x}_i$ for simplicity. We can divide equation (8.24) by $\sqrt{h_i}$ to get

$$\begin{aligned} y_i/\sqrt{h_i} = \beta_0/\sqrt{h_i} + \beta_1(x_{i1}/\sqrt{h_i}) + \beta_2(x_{i2}/\sqrt{h_i}) + \cdots \\ + \beta_k(x_{ik}/\sqrt{h_i}) + (u_i/\sqrt{h_i}) \end{aligned} \qquad [8.25]$$

or

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + u_i^*, \qquad [8.26]$$

where $x_{i0}^* = 1/\sqrt{h_i}$ and the other starred variables denote the corresponding original variables divided by $\sqrt{h_i}$.

Equation (8.26) looks a little peculiar, but the important thing to remember is that we derived it so we could obtain estimators of the β_j that have better efficiency properties than OLS. The intercept β_0 in the original equation (8.24) is now multiplying the variable $x_{i0}^* = 1/\sqrt{h_i}$. Each slope parameter in β_j multiplies a new variable that rarely has a useful interpretation. This should not cause problems if we recall that, for interpreting the parameters and the model, we always want to return to the original equation (8.24).

In the preceding savings example, the transformed equation looks like

$$sav_i/\sqrt{inc_i} = \beta_0(1/\sqrt{inc_i}) + \beta_1\sqrt{inc_i} + u_i^*,$$

where we use the fact that $inc_i/\sqrt{inc_i} = \sqrt{inc_i}$. Nevertheless, β_1 is the marginal propensity to save out of income, an interpretation we obtain from equation (8.22).

Equation (8.26) is linear in its parameters (so it satisfies MLR.1), and the random sampling assumption has not changed. Further, u_i^* has a zero mean and a constant variance (σ^2), conditional on $\mathbf{x}_i^*$. This means that if the original equation satisfies the first four Gauss-Markov assumptions, then the transformed equation (8.26) satisfies all five Gauss-Markov assumptions. Also, if u_i has a normal distribution, then u_i^* has a normal distribution with variance σ^2. Therefore, the transformed equation satisfies the classical linear model assumptions (MLR.1 through MLR.6) if the original model does so except for the homoskedasticity assumption.

Because we know that OLS has appealing properties (is BLUE, for example) under the Gauss-Markov assumptions, the discussion in the previous paragraph suggests estimating the parameters in equation (8.26) by ordinary least squares. These estimators, $\beta_0^*, \beta_1^*, \ldots, \beta_k^*$, will be different from the OLS estimators in the original equation. The β_j^* are examples of **generalized least squares (GLS) estimators**. In this case, the GLS estimators are used to account for heteroskedasticity in the errors. We will encounter other GLS estimators in Chapter 12.

Because equation (8.26) satisfies all of the ideal assumptions, standard errors, t statistics, and F statistics can all be obtained from regressions using the transformed variables. The sum of squared residuals from (8.26) divided by the degrees of freedom is an unbiased estimator of σ^2. Further, the GLS estimators, because they are the best linear unbiased estimators of the β_j, are necessarily more

efficient than the OLS estimators $\hat{\beta}_j$ obtained from the untransformed equation. Essentially, after we have transformed the variables, we simply use standard OLS analysis. But we must remember to interpret the estimates in light of the original equation.

The GLS estimators for correcting heteroskedasticity are called **weighted least squares (WLS) estimators**. This name comes from the fact that the β_j^* minimize the *weighted* sum of squared residuals, where each squared residual is weighted by $1/h_i$. The idea is that less weight is given to observations with a higher error variance; OLS gives each observation the same weight because it is best when the error variance is identical for all partitions of the population. Mathematically, the WLS estimators are the values of the b_j that make

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik})^2 / h_i \qquad [8.27]$$

as small as possible. Bringing the square root of $1/h_i$ inside the squared residual shows that the weighted sum of squared residuals is identical to the sum of squared residuals in the transformed variables:

$$\sum_{i=1}^{n} (y_i^* - b_0 x_{i0}^* - b_1 x_{i1}^* - b_2 x_{i2}^* - \cdots - b_k x_{ik}^*)^2.$$

Because OLS minimizes the sum of squared residuals (regardless of the definitions of the dependent variable and independent variable), it follows that the WLS estimators that minimize (8.27) are simply the OLS estimators from (8.26). Note carefully that the squared residuals in (8.27) are weighted by $1/h_i$, whereas the transformed variables in (8.26) are weighted by $1/\sqrt{h_i}$.

A weighted least squares estimator can be defined for any set of positive weights. Ordinary least squares is the special case that gives equal weight to all observations. The efficient procedure, GLS, weights each squared residual by the *inverse* of the conditional variance of u_i given $\mathbf{x}_i$.

Obtaining the transformed variables in equation (8.25) in order to manually perform weighted least squares can be tedious, and the chance of making mistakes is nontrivial. Fortunately, most modern regression packages have a feature for computing weighted least squares. Typically, along with the dependent and independent variables in the original model, we just specify the weighting function, $1/h_i$, appearing in (8.27). That is, we specify weights proportional to the inverse of the variance. In addition to making mistakes less likely, this forces us to interpret weighted least squares estimates in the original model. In fact, we can write out the estimated equation in the usual way. The estimates and standard errors will be different from OLS, but the way we *interpret* those estimates, standard errors, and test statistics is the same.

Econometrics packages that have a built-in WLS option will report an R-squared (and adjusted R-squared) along with WLS estimates and standard errors. Typically, the WLS R-squared is obtained from the weighted SSR, obtained from minimizing equation (8.27), and a weighted total sum of squares (SST), obtained by using the same weights but setting all of the slope coefficients in equation (8.27), $b_1, b_2, \ldots, b_k$, to zero. As a goodness-of-fit measure, this R-squared is not especially useful, as it effectively measures explained variation in y_i^* rather than y_i. Nevertheless, the WLS R-squareds computed as just described are appropriate for computing F statistics for exclusion restrictions (provided we have properly specified the variance function). As in the case of OLS, the SST terms cancel, and so we obtain the F statistic based on the weighted SSR.

The R-squared from running the OLS regression in equation (8.26) is even less useful as a goodness-of-fit measure, as the computation of SST would make little sense: one would necessarily exclude an intercept from the regression, in which case regression packages typically compute the SST without properly centering the y_i^*. This is another reason for using a WLS option that is pre-programmed in a regression package because at least the reported R-squared properly compares the model with all of the independent variables to a model with only an intercept. Because the SST cancels out when testing exclusion restrictions, improperly computing SST does not affect the R-squared form of the F statistic. Nevertheless, computing such an R-squared tempts one to think the equation fits better than it does.

EXAMPLE 8.6 Financial Wealth Equation

We now estimate equations that explain net total financial wealth (*nettfa*, measured in $1,000s) in terms of income (*inc*, also measured in $1,000s) and some other variables, including age, gender, and an indicator for whether the person is eligible for a 401(k) pension plan. We use the data on single people ($fsize = 1$) in 401KSUBS. In Computer Exercise C12 in Chapter 6, it was found that a specific quadratic function in *age*, namely $(age - 25)^2$, fit the data just as well as an unrestricted quadratic. Plus, the restricted form gives a simplified interpretation because the minimum age in the sample is 25: *nettfa* is an increasing function of *age* after $age = 25$.

The results are reported in Table 8.1. Because we suspect heteroskedasticity, we report the heteroskedasticity-robust standard errors for OLS. The weighted least squares estimates, and their standard errors, are obtained under the assumption $\text{Var}(u|inc) = \sigma^2 inc$.

Without controlling for other factors, another dollar of income is estimated to increase *nettfa* by about 82¢ when OLS is used; the WLS estimate is smaller, about 79¢. The difference is not large; we certainly do not expect them to be identical. The WLS coefficient does have a smaller standard error than OLS, almost 40% smaller, provided we assume the model $\text{Var}(nettfa|inc) = \sigma^2 inc$ is correct.

Adding the other controls reduced the *inc* coefficient somewhat, with the OLS estimate still larger than the WLS estimate. Again, the WLS estimate of β_{inc} is more precise. Age has an increasing effect starting at $age = 25$, with the OLS estimate showing a larger effect. The WLS estimate of β_{age} is more precise in this case. Gender does not have a statistically significant effect on *nettfa*, but being eligible for a 401(k) plan does: the OLS estimate is that those eligible, holding fixed income, age, and gender, have net total financial assets about $6,890 higher. The WLS estimate is substantially below the OLS estimate and suggests a misspecification of the functional form in the mean equation. (One possibility is to interact *e401k* and *inc*; see Computer Exercise C11.)

GOING FURTHER 8.3

Using the OLS residuals obtained from the OLS regression reported in column (1) of Table 8.1, the regression of $\hat{u}^2$ on *inc* yields a *t* statistic of 2.96. Does it appear we should worry about heteroskedasticity in the financial wealth equation?

Using WLS, the *F* statistic for joint significance of $(age - 25)^2$, *male*, and *e401k* is about 30.8 if we use the *R*-squareds reported in Table 8.1. With 3 and 2,012 degrees of freedom, the *p*-value is zero to more than 15 decimal places; of course, this is not surprising given the very large *t* statistics for the age and 401(k) variables.

TABLE 8.1 Dependent Variable: *nettfa*

Independent Variables	(1) OLS	(2) WLS	(3) OLS	(4) WLS
inc	.821 (.104)	.787 (.063)	.771 (.100)	.740 (.064)
$(age - 25)^2$	—	—	.0251 (.0043)	.0175 (.0019)
male	—	—	2.48 (2.06)	1.84 (1.56)
e401k	—	—	6.89 (2.29)	5.19 (1.70)
intercept	−10.57 (2.53)	−9.58 (1.65)	−20.98 (3.50)	−16.70 (1.96)
Observations	2,017	2,017	2,017	2,017
R-squared	.0827	.0709	.1279	.1115

Assuming that the error variance in the financial wealth equation has a variance proportional to income is essentially arbitrary. In fact, in most cases, our choice of weights in WLS has a degree of arbitrariness. However, there is one case in which the weights needed for WLS arise naturally from an underlying econometric model. This happens when, instead of using individual-level data, we only have averages of data across some group or geographic region. For example, suppose we are interested in determining the relationship between the amount a worker contributes to his or her 401(k) pension plan as a function of the plan generosity. Let i denote a particular firm and let e denote an employee within the firm. A simple model is

$$contrib_{i,e} = \beta_0 + \beta_1 earns_{i,e} + \beta_2 age_{i,e} + \beta_3 mrate_i + u_{i,e}, \qquad [8.28]$$

where $contrib_{i,e}$ is the annual contribution by employee e who works for firm i, $earns_{i,e}$ is annual earnings for this person, and $age_{i,e}$ is the person's age. The variable $mrate_i$ is the amount the firm puts into an employee's account for every dollar the employee contributes.

If (8.28) satisfies the Gauss-Markov assumptions, then we could estimate it, given a sample of individuals across various employers. Suppose, however, that we only have *average* values of contributions, earnings, and age by employer. In other words, individual-level data are not available. Thus, let $\overline{contrib}_i$ denote average contribution for people at firm i, and similarly for $\overline{earns}_i$ and $\overline{age}_i$. Let m_i denote the number of employees at firm i; we assume that this is a known quantity. Then, if we average equation (8.28) across all employees at firm i, we obtain the firm-level equation

$$\overline{contrib}_i = \beta_0 + \beta_1 \overline{earns}_i + \beta_2 \overline{age}_i + \beta_3 mrate_i + \overline{u}_i, \qquad [8.29]$$

where $\overline{u}_i = m_i^{-1} \sum_{e=1}^{mi} u_{i,e}$ is the average error across all employees in firm i. If we have n firms in our sample, then (8.29) is just a standard multiple linear regression model that can be estimated by OLS. The estimators are unbiased if the original model (8.28) satisfies the Gauss-Markov assumptions and the individual errors $u_{i,e}$ are independent of the firm's size, m_i [because then the expected value of $\overline{u}_i$, given the explanatory variables in (8.29), is zero].

If the individual-level equation (8.28) satisfies the homoskedasticity assumption, and the errors within firm i are uncorrelated across employees, then we can show that the firm-level equation (8.29) has a particular kind of heteroskedasticity. Specifically, if $\text{Var}(u_{i,e}) = \sigma^2$ for all i and e, and $\text{Cov}(u_{i,e}, u_{i,g}) = 0$ for every pair of employees $e \neq g$ within firm i, then $\text{Var}(\overline{u}_i) = \sigma^2/m_i$; this is just the usual formula for the variance of an average of uncorrelated random variables with common variance. In other words, the variance of the error term $\overline{u}_i$ decreases with firm size. In this case, $h_i = 1/m_i$, and so the most efficient procedure is weighted least squares, with weights equal to the number of employees at the firm $(1/h_i = m_i)$. This ensures that larger firms receive more weight. This gives us an efficient way of estimating the parameters in the individual-level model when we only have averages at the firm level.

A similar weighting arises when we are using per capita data at the city, county, state, or country level. If the individual-level equation satisfies the Gauss-Markov assumptions, then the error in the per capita equation has a variance proportional to one over the size of the population. Therefore, weighted least squares with weights equal to the population is appropriate. For example, suppose we have city-level data on per capita beer consumption (in ounces), the percentage of people in the population over 21 years old, average adult education levels, average income levels, and the city price of beer. Then, the city-level model

$$beerpc = \beta_0 + \beta_1 perc21 + \beta_2 avgeduc + \beta_3 incpc + \beta_4 price + u$$

can be estimated by weighted least squares, with the weights being the city population.

The advantage of weighting by firm size, city population, and so on relies on the underlying individual equation being homoskedastic. If heteroskedasticity exists at the individual level, then the proper weighting depends on the form of heteroskedasticity. Further, if there is correlation across errors within a group (say, firm), then $\text{Var}(\overline{u}_i) \neq \sigma^2/m_i$; see Problem 7. Uncertainty about the form of $\text{Var}(\overline{u}_i)$ in equations such as (8.29) is why more and more researchers simply use OLS and compute

robust standard errors and test statistics when estimating models using per capita data. An alternative is to weight by group size but to report the heteroskedasticity-robust statistics in the WLS estimation. This ensures that, while the estimation is efficient if the individual-level model satisfies the Gauss-Markov assumptions, heteroskedasticity at the individual level or within-group correlation are accounted for through robust inference.

8-4b The Heteroskedasticity Function Must Be Estimated: Feasible GLS

In the previous subsection, we saw some examples of where the heteroskedasticity is known up to a multiplicative form. In most cases, the exact form of heteroskedasticity is not obvious. In other words, it is difficult to find the function $h(\mathbf{x}_i)$ of the previous section. Nevertheless, in many cases we can model the function h and use the data to estimate the unknown parameters in this model. This results in an estimate of each h_i, denoted as $\hat{h}_i$. Using $\hat{h}_i$ instead of h_i in the GLS transformation yields an estimator called the **feasible GLS (FGLS) estimator**. Feasible GLS is sometimes called *estimated GLS*, or EGLS.

There are many ways to model heteroskedasticity, but we will study one particular, fairly flexible approach. Assume that

$$\text{Var}(u|\mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k), \tag{8.30}$$

where $x_1, x_2, \ldots, x_k$ are the independent variables appearing in the regression model [see equation (8.1)], and the δ_j are unknown parameters. Other functions of the x_j can appear, but we will focus primarily on (8.30). In the notation of the previous subsection, $h(\mathbf{x}) = \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k)$.

You may wonder why we have used the exponential function in (8.30). After all, when *testing* for heteroskedasticity using the Breusch-Pagan test, we assumed that heteroskedasticity was a linear function of the x_j. Linear alternatives such as (8.12) are fine when testing for heteroskedasticity, but they can be problematic when correcting for heteroskedasticity using weighted least squares. We have encountered the reason for this problem before: linear models do not ensure that predicted values are positive, and our estimated variances must be positive in order to perform WLS.

If the parameters δ_j were known, then we would just apply WLS, as in the previous subsection. This is not very realistic. It is better to use the data to estimate these parameters, and then to use these estimates to construct weights. How can we estimate the δ_j? Essentially, we will transform this equation into a linear form that, with slight modification, can be estimated by OLS.

Under assumption (8.30), we can write

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k) v,$$

where v has a mean equal to unity, conditional on $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. If we assume that v is actually independent of $\mathbf{x}$, we can write

$$\log(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + e, \tag{8.31}$$

where e has a zero mean and is independent of $\mathbf{x}$; the intercept in this equation is different from δ_0, but this is not important in implementing WLS. The dependent variable is the log of the squared error. Because (8.31) satisfies the Gauss-Markov assumptions, we can get unbiased estimators of the δ_j by using OLS.

As usual, we must replace the unobserved u with the OLS residuals. Therefore, we run the regression of

$$\log(\hat{u}^2) \text{ on } x_1, x_2, \ldots, x_k. \tag{8.32}$$

Actually, what we need from this regression are the fitted values; call these $\hat{g}_i$. Then, the estimates of h_i are simply

$$\hat{h}_i = \exp(\hat{g}_i). \tag{8.33}$$

We now use WLS with weights $1/\hat{h}_i$ in place of $1/h_i$ in equation (8.27). We summarize the steps.

A Feasible GLS Procedure to Correct for Heteroskedasticity:

1. Run the regression of y on $x_1, x_2, \ldots, x_k$ and obtain the residuals, $\hat{u}$.
2. Create $\log(\hat{u}^2)$ by first squaring the OLS residuals and then taking the natural log.
3. Run the regression in equation (8.32) and obtain the fitted values, $\hat{g}$.
4. Exponentiate the fitted values from (8.32): $\hat{h} = \exp(\hat{g})$.
5. Estimate the equation

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

by WLS, using weights $1/\hat{h}$. In other words, we replace h_i with $\hat{h}_i$ in equation (8.27). Remember, the *squared* residual for observation i gets weighted by $1/\hat{h}_i$. If instead we first transform all variables and run OLS, each variable gets multiplied by $1/\sqrt{\hat{h}_i}$, including the intercept.

If we could use h_i rather than $\hat{h}_i$ in the WLS procedure, we know that our estimators would be unbiased; in fact, they would be the best linear unbiased estimators, assuming that we have properly modeled the heteroskedasticity. Having to estimate h_i using the same data means that the FGLS estimator is no longer unbiased (so it cannot be BLUE, either). Nevertheless, the FGLS estimator is consistent and *asymptotically* more efficient than OLS. This is difficult to show because of estimation of the variance parameters. But if we ignore this—as it turns out we may—the proof is similar to showing that OLS is efficient in the class of estimators in Theorem 5.3. At any rate, for large sample sizes, FGLS is an attractive alternative to OLS when there is evidence of heteroskedasticity that inflates the standard errors of the OLS estimates.

We must remember that the FGLS estimators are estimators of the parameters in the usual population model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u.$$

Just as the OLS estimates measure the marginal impact of each x_j on y, so do the FGLS estimates. We use the FGLS estimates in place of the OLS estimates because the FGLS estimators are more efficient and have associated test statistics with the usual t and F distributions, at least in large samples. If we have some doubt about the variance specified in equation (8.30), we can use heteroskedasticity-robust standard errors and test statistics in the transformed equation.

Another useful alternative for estimating h_i is to replace the independent variables in regression (8.32) with the OLS fitted values and their squares. In other words, obtain the $\hat{g}_i$ as the fitted values from the regression of

$$\log(\hat{u}^2) \text{ on } \hat{y}, \hat{y}^2 \qquad [8.34]$$

and then obtain the $\hat{h}_i$ exactly as in equation (8.33). This changes only step (3) in the previous procedure.

If we use regression (8.32) to estimate the variance function, you may be wondering if we can simply test for heteroskedasticity using this same regression (an F or LM test can be used). In fact, Park (1966) suggested this. Unfortunately, when compared with the tests discussed in Section 8-3, the Park test has some problems. First, the null hypothesis must be something stronger than homoskedasticity: effectively, u and $\mathbf{x}$ must be independent. This is not required in the Breusch-Pagan or White tests. Second, using the OLS residuals $\hat{u}$ in place of u in (8.32) can cause the F statistic to deviate from the F distribution, even in large sample sizes. This is not an issue in the other tests we have covered. For these reasons, the Park test is not recommended when testing for heteroskedasticity. Regression (8.32) works well for weighted least squares because we only need consistent estimators of the δ_j, and regression (8.32) certainly delivers those.

EXAMPLE 8.7 Demand for Cigarettes

We use the data in SMOKE to estimate a demand function for daily cigarette consumption. Because most people do not smoke, the dependent variable, *cigs*, is zero for most observations. A linear model is not ideal because it can result in negative predicted values. Nevertheless, we can still learn something about the determinants of cigarette smoking by using a linear model.

The equation estimated by ordinary least squares, with the usual OLS standard errors in parentheses, is

$$\begin{aligned}\widehat{cigs} &= -3.64 + .880\log(income) - .751\log(cigpric)\\ &\quad\ (24.08)\ \ (.728)\qquad\quad (5.773)\\ &\quad -.501\,educ + .771\,age - .0090\,age^2 - 2.83\,restaurn\\ &\quad\ (.167)\qquad (.160)\qquad (.0017)\qquad (1.11)\\ n &= 807,\ R^2 = .0526,\end{aligned} \tag{8.35}$$

where

cigs = number of cigarettes smoked per day.
income = annual income.
cigpric = the per-pack price of cigarettes (in cents).
educ = years of schooling.
age = age measured in years.
restaurn = a binary indicator equal to unity if the person resides in a state with restaurant smoking restrictions.

Because we are also going to do weighted least squares, we do not report the heteroskedasticity-robust standard errors for OLS. (Incidentally, 13 out of the 807 fitted values are less than zero; this is less than 2% of the sample and is not a major cause for concern.)

Neither income nor cigarette price is statistically significant in (8.35), and their effects are not practically large. For example, if income increases by 10%, *cigs* is predicted to increase by $(.880/100)(10) = .088$, or less than one-tenth of a cigarette per day. The magnitude of the price effect is similar.

Each year of education reduces the average cigarettes smoked per day by one-half of a cigarette, and the effect is statistically significant. Cigarette smoking is also related to age, in a quadratic fashion. Smoking increases with age up until $age = .771/[2(.009)] \approx 42.83$, and then smoking decreases with age. Both terms in the quadratic are statistically significant. The presence of a restriction on smoking in restaurants decreases cigarette smoking by almost three cigarettes per day, on average.

Do the errors underlying equation (8.35) contain heteroskedasticity? The Breusch-Pagan regression of the squared OLS residuals on the independent variables in (8.35) [see equation (8.14)] produces $R^2_{\hat{u}2} = .040$. This small R-squared may seem to indicate no heteroskedasticity, but we must remember to compute either the F or LM statistic. If the sample size is large, a seemingly small $R^2_{\hat{u}2}$ can result in a very strong rejection of homoskedasticity. The LM statistic is $LM = 807(.040) = 32.28$, and this is the outcome of a χ^2_6 random variable. The p-value is less than .000015, which is very strong evidence of heteroskedasticity.

Therefore, we estimate the equation using the feasible GLS procedure based on equation (8.32). The weighted least squares estimates are

$$\begin{aligned}\widehat{cigs} &= 5.64 + 1.30\log(income) - 2.94\log(cigpric)\\ &\quad\ (17.80)\ \ (.44)\qquad\quad (4.46)\\ &\quad -.463\,educ + .482\,age - .0056\,age^2 - 3.46\,restaurn\\ &\quad\ (.120)\qquad (.097)\qquad (.0009)\qquad (.80)\\ n &= 807,\ R^2 = .1134.\end{aligned} \tag{8.36}$$

The income effect is now statistically significant and larger in magnitude. The price effect is also notably bigger, but it is still statistically insignificant. [One reason for this is that *cigpric* varies only across states in the sample, and so there is much less variation in log(*cigpric*) than in log(*income*), *educ*, and *age*.]

The estimates on the other variables have, naturally, changed somewhat, but the basic story is still the same. Cigarette smoking is negatively related to schooling, has a quadratic relationship with *age*, and is negatively affected by restaurant smoking restrictions.

We must be a little careful in computing F statistics for testing multiple hypotheses after estimation by WLS. (This is true whether the sum of squared residuals or R-squared form of the F statistic is used.) It is important that the same weights be used to estimate the unrestricted and restricted models. We should first estimate the unrestricted model by OLS. Once we have obtained the weights, we can use them to estimate the restricted model as well. The F statistic can be computed as usual. Fortunately, many regression packages have a simple command for testing joint restrictions after WLS estimation, so we need not perform the restricted regression ourselves.

Example 8.7 hints at an issue that sometimes arises in applications of weighted least squares: the OLS and WLS estimates can be substantially different. This is not such a big problem in the demand for cigarettes equation because all the coefficients maintain the same signs, and the biggest changes are on variables that were statistically insignificant when the equation was estimated by OLS. The OLS and WLS estimates will always differ due to sampling error. The issue is whether their difference is enough to change important conclusions.

GOING FURTHER 8.4

Let $\hat{u}_i$ be the WLS residuals from (8.36), which are not weighted, and let $\widehat{cigs}_i$ be the fitted values. (These are obtained using the same formulas as OLS; they differ because of different estimates of the β_j.) One way to determine whether heteroskedasticity has been eliminated is to use the $\hat{u}_i^2/\hat{h}_i = (\hat{u}_i/\sqrt{\hat{h}_i})^2$ in a test for heteroskedasticity. [If $h_i = \text{Var}(u_i|\mathbf{x}_i)$, then the transformed residuals should have little evidence of heteroskedasticity.] There are many possibilities, but one—based on White's test in the transformed equation—is to regress $\hat{u}_i^2/\hat{h}_i$ on $\widehat{cigs}_i/\sqrt{\hat{h}_i}$ and $\widehat{cigs}_i^2/\hat{h}_i$ (including an intercept). The joint F statistic when we use SMOKE is 11.15. Does it appear that our correction for heteroskedasticity has actually eliminated the heteroskedasticity?

If OLS and WLS produce statistically significant estimates that differ in sign—for example, the OLS price elasticity is positive and significant, while the WLS price elasticity is negative and significant—or the difference in magnitudes of the estimates is practically large, we should be suspicious. Typically, this indicates that one of the *other* Gauss-Markov assumptions is false, particularly the zero conditional mean assumption on the error (MLR.4). If $\text{E}(y|\mathbf{x}) \neq \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, then OLS and WLS have different expected values and probability limits. For WLS to be consistent for the β_j, it is not enough for u to be uncorrelated with each x_j; we need the stronger assumption MLR.4 in the linear model MLR.1. Therefore, a significant difference between OLS and WLS can indicate a functional form misspecification in $\text{E}(y|\mathbf{x})$. The *Hausman test* [Hausman (1978)] can be used to formally compare the OLS and WLS estimates to see if they differ by more than sampling error suggests they should, but this test is beyond the scope of this text. In many cases, an informal "eyeballing" of the estimates is sufficient to detect a problem.

8-4c What If the Assumed Heteroskedasticity Function Is Wrong?

We just noted that if OLS and WLS produce very different estimates, it is likely that the conditional mean $\text{E}(y|\mathbf{x})$ is misspecified. What are the properties of WLS if the variance function we use is misspecified in the sense that $\text{Var}(y|\mathbf{x}) \neq \sigma^2 h(\mathbf{x})$ for our chosen function $h(\mathbf{x})$? The most important issue

is whether misspecification of $h(\mathbf{x})$ causes bias or inconsistency in the WLS estimator. Fortunately, the answer is no, at least under MLR.4. Recall that, if $\mathrm{E}(u|\mathbf{x}) = 0$, then any function of $\mathbf{x}$ is uncorrelated with u, and so the weighted error, $u/\sqrt{h(\mathbf{x})}$, is uncorrelated with the weighted regressors, $x_j/\sqrt{h(\mathbf{x})}$, for *any* function $h(\mathbf{x})$ that is always positive. This is why, as we just discussed, we can take large differences between the OLS and WLS estimators as indicative of functional form misspecification. If we estimate parameters in the function, say $h(\mathbf{x}, \hat{\delta})$, then we can no longer claim that WLS is unbiased, but it will generally be consistent (whether or not the variance function is correctly specified).

If WLS is at least consistent under MLR.1 to MLR.4, what are the consequences of using WLS with a misspecified variance function? There are two. The first, which is very important, is that the usual WLS standard errors and test statistics, computed under the assumption that $\mathrm{Var}(y|\mathbf{x}) = \sigma^2 h(\mathbf{x})$, are no longer valid, even in large samples. For example, the WLS estimates and standard errors in column (4) of Table 8.1 assume that $\mathrm{Var}(nettfa|inc, age, male, e401k) = \mathrm{Var}(nettfa|inc) = \sigma^2 inc$; so we are assuming not only that the variance depends just on income, but also that it is a linear function of income. If this assumption is false, the standard errors (and any statistics we obtain using those standard errors) are not valid. Fortunately, there is an easy fix: just as we can obtain standard errors for the OLS estimates that are robust to arbitrary heteroskedasticity, we can obtain standard errors for WLS that allow the variance function to be arbitrarily misspecified. It is easy to see why this works. Write the transformed equation as

$$y_i/\sqrt{h_i} = \beta_0(1/\sqrt{h_i}) + \beta_1(x_{i1}/\sqrt{h_i}) + \cdots + \beta_k(x_{ik}/\sqrt{h_i}) + u_i/\sqrt{h_i}.$$

Now, if $\mathrm{Var}(u_i|\mathbf{x}_i) \neq \sigma^2 h_i$, then the weighted error $u_i/\sqrt{h_i}$ is heteroskedastic. So we can just apply the usual heteroskedasticity-robust standard errors after estimating this equation by OLS—which, remember, is identical to WLS.

To see how robust inference with WLS works in practice, column (1) of Table 8.2 reproduces the last column of Table 8.1, and column (2) contains standard errors robust to $\mathrm{Var}(u_i|\mathbf{x}_i) \neq \sigma^2 inc_i$.

The standard errors in column (2) allow the variance function to be misspecified. We see that, for the income and age variables, the robust standard errors are somewhat above the usual WLS standard errors—certainly by enough to stretch the confidence intervals. On the other hand, the robust standard errors for *male* and *e401k* are actually smaller than those that assume a correct variance function. We saw this could happen with the heteroskedasticity-robust standard errors for OLS, too.

Even if we use flexible forms of variance functions, such as that in (8.30), there is no guarantee that we have the correct model. While exponential heteroskedasticity is appealing and reasonably flexible, it is, after all, just a model. Therefore, it is always a good idea to compute fully robust standard errors and test statistics after WLS estimation.

TABLE 8.2 WLS Estimation of the *nettfa* Equation

Independent Variables	With Nonrobust Standard Errors	With Robust Standard Errors
inc	.740 (.064)	.740 (.075)
$(age - 25)^2$	.0175 (.0019)	.0175 (.0026)
male	1.84 (1.56)	1.84 (1.31)
e401k	5.19 (1.70)	5.19 (1.57)
intercept	−16.70 (1.96)	−16.70 (2.24)
Observations	2,017	2,017
R-squared	.1115	.1115

A modern criticism of WLS is that if the variance function is misspecified, it is not guaranteed to be more efficient than OLS. In fact, that is the case: if $\text{Var}(y|\mathbf{x})$ is neither constant nor equal to $\sigma^2 h(\mathbf{x})$, where $h(\mathbf{x})$ is the proposed model of heteroskedasticity, then we cannot rank OLS and WLS in terms of variances (or asymptotic variances when the variance parameters must be estimated). However, this theoretically correct criticism misses an important practical point. Namely, in cases of strong heteroskedasticity, it is often better to use a wrong form of heteroskedasticity and apply WLS than to ignore heteroskedasticity altogether in estimation and use OLS. Models such as (8.30) can well approximate a variety of heteroskedasticity functions and may produce estimators with smaller (asymptotic) variances. Even in Example 8.6, where the form of heteroskedasticity was assumed to have the simple form $\text{Var}(nettfa|\mathbf{x}) = \sigma^2 inc$, the fully robust standard errors for WLS are well below the fully robust standard errors for OLS. (Comparing robust standard errors for the two estimators puts them on equal footing: we assume neither homoskedasticity nor that the variance has the form $\sigma^2 inc$.) For example, the robust standard error for the WLS estimator of β_{inc} is about .075, which is 25% lower than the robust standard error for OLS (about .100). For the coefficient on $(age - 25)^2$, the robust standard error of WLS is about .0026, almost 40% below the robust standard error for OLS (about .0043).

8-4d Prediction and Prediction Intervals with Heteroskedasticity

If we start with the standard linear model under MLR.1 to MLR.4, but allow for heteroskedasticity of the form $\text{Var}(y|\mathbf{x}) = \sigma^2 h(\mathbf{x})$ [see equation (8.21)], the presence of heteroskedasticity affects the point prediction of y only insofar as it affects estimation of the β_j. Of course, it is natural to use WLS on a sample of size n to obtain the $\hat{\beta}_j$. Our prediction of an unobserved outcome, y^0, given known values of the explanatory variables $\mathbf{x}^0$, has the same form as in Section 6-4: $\hat{y}^0 = \hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}}$. This makes sense: once we know $\text{E}(y|\mathbf{x})$, we base our prediction on it; the structure of $\text{Var}(y|\mathbf{x})$ plays no direct role.

On the other hand, prediction *intervals* do depend directly on the nature of $\text{Var}(y|\mathbf{x})$. Recall in Section 6-4 that we constructed a prediction interval under the classical linear model assumptions. Suppose now that all the CLM assumptions hold except that (8.21) replaces the homoskedasticity assumption, MLR.5. We know that the WLS estimators are BLUE and, because of normality, have (conditional) normal distributions. We can obtain $\text{se}(\hat{y}^0)$ using the same method in Section 6-4, except that now we use WLS. [A simple approach is to write $y_i = \theta_0 + \beta_1(x_{i1} - x_1^0) + \cdots + \beta_k(x_{ik} - x_k^0) + u_i$, where the x_j^0 are the values of the explanatory variables for which we want a predicted value of y. We can estimate this equation by WLS and then obtain $\hat{y}^0 = \hat{\theta}_0$ and $\text{se}(\hat{y}^0) = \text{se}(\hat{\theta}_0)$.] We also need to estimate the standard deviation of u^0, the unobserved part of y^0. But $\text{Var}(u^0|\mathbf{x} = \mathbf{x}^0) = \sigma^2 h(\mathbf{x}^0)$, and so $\text{se}(u^0) = \hat{\sigma}\sqrt{h(\mathbf{x}^0)}$, where $\hat{\sigma}$ is the standard error of the regression from the WLS estimation. Therefore, a 95% prediction interval is

$$\hat{y}^0 \pm t_{.025} \cdot \text{se}(\hat{e}^0), \qquad [8.37]$$

where $\text{se}(\hat{e}^0) = \{[\text{se}(\hat{y}^0)]^2 + \hat{\sigma}^2 h(\mathbf{x}^0)\}^{1/2}$.

This interval is exact only if we do not have to estimate the variance function. If we estimate parameters, as in model (8.30), then we cannot obtain an exact interval. In fact, accounting for the estimation error in the $\hat{\beta}_j$ and the $\hat{\delta}_j$ (the variance parameters) becomes very difficult. We saw two examples in Section 6-4 where the estimation error in the parameters was swamped by the variation in the unobservables, u^0. Therefore, we might still use equation (8.37) with $h(\mathbf{x}^0)$ simply replaced by $\hat{h}(\mathbf{x}^0)$. In fact, if we are to ignore the parameter estimation error entirely, we can drop $\text{se}(\hat{y}^0)$ from $\text{se}(\hat{e}^0)$. [Remember, $\text{se}(\hat{y}^0)$ converges to zero at the rate $1/\sqrt{n}$, while $\text{se}(\hat{u}^0)$ is roughly constant.]

We can also obtain a prediction for y in the model

$$\log(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \qquad [8.38]$$

where u is heteroskedastic. We assume that u has a conditional normal distribution with a specific form of heteroskedasticity. We assume the exponential form in equation (8.30), but add the normality assumption:

$$u|x_1, x_2, \ldots, x_k \sim \text{Normal}[0, \exp(\delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k)]. \qquad [8.39]$$

As a notational shorthand, write the variance function as $\exp(\delta_0 + \mathbf{x}\boldsymbol{\delta})$. Then, because $\log(y)$ given $\mathbf{x}$ has a normal distribution with mean $\beta_0 + \mathbf{x}\boldsymbol{\beta}$ and variance $\exp(\delta_0 + \mathbf{x}\boldsymbol{\delta})$, it follows that

$$\text{E}(y|\mathbf{x}) = \exp(\beta_0 + \mathbf{x}\boldsymbol{\beta} + \exp(\delta_0 + \mathbf{x}\boldsymbol{\delta})/2). \qquad [8.40]$$

Now, we estimate the β_j and δ_j using WLS estimation of (8.38). That is, after using OLS to obtain the residuals, run the regression in (8.32) to obtain fitted values,

$$\hat{g}_i = \hat{\alpha}_0 + \hat{\delta}_1 x_{i1} + \cdots + \hat{\delta}_k x_{ik}, \qquad [8.41]$$

and then compute the $\hat{h}_i$ as in (8.33). Using these $\hat{h}_i$, obtain the WLS estimates, $\hat{\beta}_j$, and also compute $\hat{\sigma}^2$ from the weighted squared residuals. Now, compared with the original model for $\text{Var}(u|\mathbf{x})$, $\delta_0 = \alpha_0 + \log(\sigma^2)$, and so $\text{Var}(u|\mathbf{x}) = \sigma^2 \exp(\alpha_0 + \delta_1 x_1 + \cdots + \delta_k x_k)$. Therefore, the estimated variance is $\hat{\sigma}^2 \exp(\hat{g}_i) = \hat{\sigma}^2 \hat{h}_i$, and the fitted value for y_i is

$$\hat{y}_i = \exp(\widehat{logy_i} + \hat{\sigma}^2 \hat{h}_i/2). \qquad [8.42]$$

We can use these fitted values to obtain an R-squared measure, as described in Section 6-4: use the squared correlation coefficient between y_i and $\hat{y}_i$.

For any values of the explanatory variables $\mathbf{x}^0$, we can estimate $\text{E}(y|\mathbf{x} = \mathbf{x}^0)$ as

$$\hat{\text{E}}(y|\mathbf{x} = \mathbf{x}^0) = \exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}} + \hat{\sigma}^2 \exp(\hat{\alpha}_0 + \mathbf{x}^0\hat{\boldsymbol{\delta}})/2), \qquad [8.43]$$

where

$\hat{\beta}_j$ = the WLS estimates.
$\hat{\alpha}_0$ = the intercept in (8.41).
$\hat{\delta}_j$ = the slopes from the same regression.
$\hat{\sigma}^2$ is obtained from the WLS estimation.

Obtaining a proper standard error for the prediction in (8.42) is very complicated analytically, but, as in Section 6-4, it would be fairly easy to obtain a standard error using a resampling method such as the bootstrap described in Appendix 6A.

Obtaining a prediction interval is more of a challenge when we estimate a model for heteroskedasticity, and a full treatment is complicated. Nevertheless, we saw in Section 6-4 two examples where the error variance swamps the estimation error, and we would make only a small mistake by ignoring the estimation error in all parameters. Using arguments similar to those in Section 6-4, an approximate 95% prediction interval (for large sample sizes) is $\exp[-1.96 \cdot \hat{\sigma}\sqrt{\hat{h}(\mathbf{x}^0)}]\exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$ to $\exp[1.96 \cdot \hat{\sigma}\sqrt{\hat{h}(\mathbf{x}^0)}]\exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$, where $\hat{h}(\mathbf{x}^0)$ is the estimated variance function evaluated at $\mathbf{x}^0$, $\hat{h}(\mathbf{x}^0) = \exp(\hat{\alpha}_0 + \hat{\delta}_1 x_1^0 + \cdots + \hat{\delta}_k x_k^0)$. As in Section 6-4, we obtain this approximate interval by simply exponentiating the endpoints.

8-5 The Linear Probability Model Revisited

As we saw in Section 7-5, when the dependent variable y is a binary variable, the model must contain heteroskedasticity, unless all of the slope parameters are zero. We are now in a position to deal with this problem.

The simplest way to deal with heteroskedasticity in the linear probability model is to continue to use OLS estimation, but to also compute robust standard errors in test statistics. This ignores the fact that we actually know the form of heteroskedasticity for the LPM. Nevertheless, OLS estimation of the LPM is simple and often produces satisfactory results.

EXAMPLE 8.8 **Labor Force Participation of Married Women**

In the labor force participation example in Section 7-5 [see equation (7.29)], we reported the usual OLS standard errors. Now, we compute the heteroskedasticity-robust standard errors as well. These are reported in brackets below the usual standard errors:

$$\begin{aligned}\widehat{inlf} &= .586 - .0034\,nwifeinc + .038\,educ + .039\,exper \\ &\quad (.154)\ (.0014)\qquad (.007)\qquad (.006) \\ &\quad [.151]\ [.0015]\qquad [.007]\qquad [.006] \\ &\quad -.00060\,exper^2 - .016\,age - .262\,kidslt6 + .0130\,kidsge6 \qquad [8.44] \\ &\quad (.00018)\qquad (.002)\qquad (.034)\qquad (.0132) \\ &\quad [.00019]\qquad [.002]\qquad [.032]\qquad [.0135] \\ n &= 753,\ R^2 = .264.\end{aligned}$$

Several of the robust and OLS standard errors are the same to the reported degree of precision; in all cases, the differences are practically very small. Therefore, while heteroskedasticity is a problem in theory, it is not in practice, at least not for this example. It often turns out that the usual OLS standard errors and test statistics are similar to their heteroskedasticity-robust counterparts. Furthermore, it requires a minimal effort to compute both.

Generally, the OLS estimators are inefficient in the LPM. Recall that the conditional variance of y in the LPM is

$$\text{Var}(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})], \qquad [8.45]$$

where

$$p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad [8.46]$$

is the response probability (probability of success, $y = 1$). It seems natural to use weighted least squares, but there are a couple of hitches. The probability $p(\mathbf{x})$ clearly depends on the unknown population parameters, β_j. Nevertheless, we do have unbiased estimators of these parameters, namely the OLS estimators. When the OLS estimators are plugged into equation (8.46), we obtain the OLS fitted values. Thus, for each observation i, $\text{Var}(y_i|\mathbf{x}_i)$ is estimated by

$$\hat{h}_i = \hat{y}_i(1 - \hat{y}_i), \qquad [8.47]$$

where $\hat{y}_i$ is the OLS fitted value for observation i. Now, we apply feasible GLS, just as in Section 8-4.

Unfortunately, being able to estimate h_i for each i does not mean that we can proceed directly with WLS estimation. The problem is one that we briefly discussed in Section 7-5: the fitted values $\hat{y}_i$ need not fall in the unit interval. If either $\hat{y}_i < 0$ or $\hat{y}_i > 1$, equation (8.47) shows that $\hat{h}_i$ will be negative. Because WLS proceeds by multiplying observation i by $1/\sqrt{\hat{h}_i}$, the method will fail if $\hat{h}_i$ is negative (or zero) for any observation. In other words, all of the weights for WLS must be positive.

In some cases, $0 < \hat{y}_i < 1$ for all i, in which case WLS can be used to estimate the LPM. In cases with many observations and small probabilities of success or failure, it is very common to find some fitted values outside the unit interval. If this happens, as it does in the labor force participation example in equation (8.44), it is easiest to abandon WLS and to report the heteroskedasticity-robust statistics. An alternative is to adjust those fitted values that are less than zero or greater than unity, and then to apply WLS. One suggestion is to set $\hat{y}_i = .01$ if $\hat{y}_i < 0$ and $\hat{y}_i = .99$ if $\hat{y}_i > 1$. Unfortunately, this requires an arbitrary choice on the part of the researcher—for example, why not use .001 and .999 as the adjusted values? If many fitted values are outside the unit interval, the adjustment to the fitted values can affect the results; in this situation, it is probably best to just use OLS.

Estimating the Linear Probability Model by Weighted Least Squares:

1. Estimate the model by OLS and obtain the fitted values, $\hat{y}$.
2. Determine whether all of the fitted values are inside the unit interval. If so, proceed to step (3). If not, some adjustment is needed to bring all fitted values into the unit interval.
3. Construct the estimated variances in equation (8.47).
4. Estimate the equation

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

by WLS, using weights $1/\hat{h}$.

EXAMPLE 8.9 **Determinants of Personal Computer Ownership**

We use the data in GPA1 to estimate the probability of owning a computer. Let *PC* denote a binary indicator equal to unity if the student owns a computer, and zero otherwise. The variable *hsGPA* is high school GPA, *ACT* is achievement test score, and *parcoll* is a binary indicator equal to unity if at least one parent attended college. (Separate college indicators for the mother and the father do not yield individually significant results, as these are pretty highly correlated.)

The equation estimated by OLS is

$$\begin{aligned} \widehat{PC} &= -.0004 + .065\, hsGPA + .0006\, ACT + .221\, parcoll \\ &\quad (.4905)\ (.137)\quad\ (.0155)\quad\ (.093) \\ &\quad [.4888]\ [.139]\quad\ [.0158]\quad\ [.087] \\ n &= 141,\ R^2 = .0415. \end{aligned} \tag{8.48}$$

Just as with Example 8.8, there are no striking differences between the usual and robust standard errors. Nevertheless, we also estimate the model by WLS. Because all of the OLS fitted values are inside the unit interval, no adjustments are needed:

$$\begin{aligned} \widehat{PC} &= .026 + .033\, hsGPA + .0043\, ACT + .215\, parcoll \\ &\quad (.477)\ (.130)\quad\ (.0155)\quad\ (.086) \\ n &= 142,\ R^2 = .0464. \end{aligned} \tag{8.49}$$

There are no important differences in the OLS and WLS estimates. The only significant explanatory variable is *parcoll*, and in both cases we estimate that the probability of PC ownership is about .22 higher if at least one parent attended college.

Summary

We began by reviewing the properties of ordinary least squares in the presence of heteroskedasticity. Heteroskedasticity does not cause bias or inconsistency in the OLS estimators, but the usual standard errors and test statistics are no longer valid. We showed how to compute heteroskedasticity-robust standard errors and *t* statistics, something that is routinely done by many regression packages. Most regression packages also compute a heteroskedasticity-robust *F*-type statistic.

We discussed two common ways to test for heteroskedasticity: the Breusch-Pagan test and a special case of the White test. Both of these statistics involve regressing the *squared* OLS residuals on either the independent variables (BP) or the fitted and squared fitted values (White). A simple *F* test is asymptotically valid; there are also Lagrange multiplier versions of the tests.

OLS is no longer the best linear unbiased estimator in the presence of heteroskedasticity. When the form of heteroskedasticity is known, GLS estimation can be used. This leads to weighted least squares as a means of obtaining the BLUE estimator. The test statistics from the WLS estimation are either exactly valid when the error term is normally distributed or asymptotically valid under nonnormality. This assumes, of course, that we have the proper model of heteroskedasticity.

More commonly, we must estimate a model for the heteroskedasticity before applying WLS. The resulting *feasible* GLS estimator is no longer unbiased, but it is consistent and asymptotically efficient. The usual statistics from the WLS regression are asymptotically valid. We discussed a method to ensure that the estimated variances are strictly positive for all observations, something needed to apply WLS.

As we discussed in Chapter 7, the linear probability model for a binary dependent variable necessarily has a heteroskedastic error term. A simple way to deal with this problem is to compute heteroskedasticity-robust statistics. Alternatively, if all the fitted values (that is, the estimated probabilities) are strictly between zero and one, weighted least squares can be used to obtain asymptotically efficient estimators.

Key Terms

Breusch-Pagan Test for Heteroskedasticity (BP Test)
Feasible GLS (FGLS) Estimator
Generalized Least Squares (GLS) Estimators
Heteroskedasticity of Unknown Form
Heteroskedasticity-Robust *F* Statistic
Heteroskedasticity-Robust *LM* Statistic
Heteroskedasticity-Robust Standard Error
Heteroskedasticity-Robust *t* Statistic
Weighted Least Squares (WLS) Estimators
White Test for Heteroskedasticity

Problems

1 Which of the following are consequences of heteroskedasticity?
(i) The OLS estimators, $\hat{\beta}_j$, are inconsistent.
(ii) The usual F statistic no longer has an F distribution.
(iii) The OLS estimators are no longer BLUE.

2 Consider a linear model to explain monthly beer consumption:

$$beer = \beta_0 + \beta_1 inc + \beta_2 price + \beta_3 educ + \beta_4 female + u$$

$$\mathrm{E}(u|inc, price, educ, female) = 0$$
$$\mathrm{Var}(u|inc, price, educ, female) = \sigma^2 inc^2.$$

Write the transformed equation that has a homoskedastic error term.

3 True or False? WLS is preferred to OLS when an important variable has been omitted from the model.

4 Using the data in GPA3, the following equation was estimated for the fall and second semester students:

$$\begin{aligned}\widehat{trmgpa} = &-2.12 + .900\, crsgpa + .193\, cumgpa + .0014\, tothrs \\ &\quad (.55)\ (.175)\quad (.064)\quad (.0012) \\ &\quad [.55]\ [.166]\quad [.074]\quad [.0012] \\ &+ .0018\, sat - .0039\, hsperc + .351\, female - .157\, season \\ &\quad (.0002)\quad (.0018)\quad (.085)\quad (.098) \\ &\quad [.0002]\quad [.0019]\quad [.079]\quad [.080] \\ n = 269,\ &R^2 = .465.\end{aligned}$$

Here, *trmgpa* is term GPA, *crsgpa* is a weighted average of overall GPA in courses taken, *cumgpa* is GPA prior to the current semester, *tothrs* is total credit hours prior to the semester, *sat* is SAT score, *hsperc* is graduating percentile in high school class, *female* is a gender dummy, and *season* is a dummy variable equal to unity if the student's sport is in season during the fall. The usual and heteroskedasticity-robust standard errors are reported in parentheses and brackets, respectively.

(i) Do the variables *crsgpa*, *cumgpa*, and *tothrs* have the expected estimated effects? Which of these variables are statistically significant at the 5% level? Does it matter which standard errors are used?

(ii) Why does the hypothesis H_0: $\beta_{crsgpa} = 1$ make sense? Test this hypothesis against the two-sided alternative at the 5% level, using both standard errors. Describe your conclusions.

(iii) Test whether there is an in-season effect on term GPA, using both standard errors. Does the significance level at which the null can be rejected depend on the standard error used?

5 The variable *smokes* is a binary variable equal to one if a person smokes, and zero otherwise. Using the data in SMOKE, we estimate a linear probability model for *smokes*:

$$\begin{aligned}\widehat{smokes} &= .656 - .069\log(cigpric) + .012\log(income) - .029\,educ\\ &\quad (.855)\ (.204)\qquad\quad (.026)\qquad\quad (.006)\\ &\quad [.856]\ [.207]\qquad\quad [.026]\qquad\quad [.006]\\ &\quad + .020\,age - .00026\,age^2 - .101\,restaurn - .026\,white\\ &\quad\ (.006)\qquad (.00006)\qquad (.039)\qquad\quad (.052)\\ &\quad\ [.005]\qquad [.00006]\qquad [.038]\qquad\quad [.050]\\ n &= 807,\ R^2 = .062.\end{aligned}$$

The variable *white* equals one if the respondent is white, and zero otherwise; the other independent variables are defined in Example 8.7. Both the usual and heteroskedasticity-robust standard errors are reported.

(i) Are there any important differences between the two sets of standard errors?

(ii) Holding other factors fixed, if education increases by four years, what happens to the estimated probability of smoking?

(iii) At what point does another year of age reduce the probability of smoking?

(iv) Interpret the coefficient on the binary variable *restaurn* (a dummy variable equal to one if the person lives in a state with restaurant smoking restrictions).

(v) Person number 206 in the data set has the following characteristics: $cigpric = 67.44$, $income = 6{,}500$, $educ = 16$, $age = 77$, $restaurn = 0$, $white = 0$, and $smokes = 0$. Compute the predicted probability of smoking for this person and comment on the result.

6 There are different ways to combine features of the Breusch-Pagan and White tests for heteroskedasticity. One possibility not covered in the text is to run the regression

$$\hat{u}_i^2 \text{ on } x_{i1}, x_{i2}, \ldots, x_{ik}, \hat{y}_i^2,\ i = 1, \ldots, n,$$

where the $\hat{u}_i$ are the OLS residuals and the $\hat{y}_i$ are the OLS fitted values. Then, we would test joint significance of $x_{i1}, x_{i2}, \ldots, x_{ik}$ and $\hat{y}_i^2$. (Of course, we always include an intercept in this regression.)

(i) What are the *df* associated with the proposed *F* test for heteroskedasticity?

(ii) Explain why the *R*-squared from the regression above will always be at least as large as the *R*-squareds for the BP regression and the special case of the White test.

(iii) Does part (ii) imply that the new test always delivers a smaller *p*-value than either the BP or special case of the White statistic? Explain.

(iv) Suppose someone suggests also adding $\hat{y}_i$ to the newly proposed test. What do you think of this idea?

7 Consider a model at the employee level,

$$y_{i,e} = \beta_0 + \beta_1 x_{i,e,1} + \beta_2 x_{i,e,2} + \cdots + \beta_k x_{i,e,k} + f_i + v_{i,e},$$

where the unobserved variable f_i is a "firm effect" to each employee at a given firm i. The error term $v_{i,e}$ is specific to employee e at firm i. The *composite error* is $u_{i,e} = f_i + v_{i,e}$, such as in equation (8.28).

(i) Assume that $\text{Var}(f_i) = \sigma_f^2$, $\text{Var}(v_{i,e}) = \sigma_v^2$, and f_i and $v_{i,e}$ are uncorrelated. Show that $\text{Var}(u_{i,e}) = \sigma_f^2 + \sigma_v^2$; call this σ^2.

(ii) Now suppose that for $e \neq g$, $v_{i,e}$ and $v_{i,g}$ are uncorrelated. Show that $\text{Cov}(u_{i,e}, u_{i,g}) = \sigma_f^2$

(iii) Let $\bar{u}_i = m_i^{-1} \sum_{\hat{e}=1}^{mi} u_{i,e}$ be the average of the composite errors within a firm. Show that $\text{Var}(\bar{u}_i) = \sigma_f^2 + \sigma_v^2/m_i$.

(iv) Discuss the relevance of part (iii) for WLS estimation using data averaged at the firm level, where the weight used for observation i is the usual firm size.

8 The following equations were estimated using the data in ECONMATH. The first equation is for men and the second is for women. The third and fourth equations combine men and women.

$$\widehat{score} = 20.52 + 13.60\, colgpa + 0.670\, act$$
$$(3.72) \quad (0.94) \quad (0.150)$$
$$n = 406.\ R^2 = .4025,\ \text{SSR} = 38{,}781.38.$$

$$\widehat{score} = 13.79 + 11.89\, colgpa + 1.03\, act$$
$$(4.11) \quad (1.09) \quad (0.18)$$
$$n = 408,\ R^2 = .3666,\ \text{SSR} = 48{,}029.82.$$

$$\widehat{score} = 15.60 + 3.17\, male + 12.82\, colgpa + 0.838\, act$$
$$(2.80) \quad (0.73) \quad (0.72) \quad (0.116)$$
$$n = 814,\ R^2 = .3946,\ \text{SSR} = 87{,}128.96.$$

$$\widehat{score} = 13.79 + 6.73\, male + 11.89\, colgpa + 1.03\, act + 1.72\, male \cdot colgpa - 0.364\, male \cdot act$$
$$(3.91) \quad (5.55) \quad (1.04) \quad (0.17) \quad (1.44) \quad (0.232)$$
$$n = 814,\ R^2 = .3968,\ \text{SSR} = 86{,}811.20.$$

(i) Compute the usual Chow statistic for testing the null hypothesis that the regression equations are the same for men and women. Find the p-value of the test.

(ii) Compute the usual Chow statistic for testing the null hypothesis that the slope coefficients are the same for men and women, and report the p-value.

(iii) Do you have enough information to compute heteroskedasticity-robust versions of the tests in (ii) and (iii)? Explain.

9 Consider the potential outcomes framework, where w is a binary treatment indicator and the potential outcomes are $y(0)$ and $y(1)$. Assume that w is randomly assigned, so that w is independent of $[y(0),y(1)]$. Let $\mu_0 = \text{E}[y(0)]$, $\mu_1 = \text{E}[y(1)]$, $\sigma_0^2 = \text{Var}[y(0)]$, and $\sigma_1^2 = \text{Var}[y(1)]$.

(i) Define the observed outcome as $y = (1 - w)y(0) + wy(1)$. Letting $\tau = \mu_1 - \mu_0$ be the average treatment effect, show you can write

$$y = \mu_0 + \tau w + (1 - w)v(0) + wv(1),$$

where $v(0) = y(0) - \mu_0$ and $v(1) = y(1) - \mu_1$.

(ii) Let $u = (1 - w)v(0) + wv(1)$ be the error term in

$$y = \mu_0 + \tau w + u$$

Show that

$$\mathrm{E}(u|w) = 0$$

What statistical properties does this finding imply about the OLS estimator of τ from the simply regression y_i on w_i for a random sample of size n? What happens as $n \to \infty$?

(iii) Show that

$$\mathrm{Var}(u|w) = \mathrm{E}(u^2|w) = (1 - w)\sigma_0^2 + w\sigma_1^2.$$

Is there generally heteroskedasticity in the error variance?

(iv) If you think $\sigma_1^2 \neq \sigma_0^2$, and $\hat{\tau}$ is the OLS estimator, how would you obtain a valid standard error for $\hat{\tau}$?

(v) After obtaining the OLS residuals, $\hat{u}_i$, $i = 1, \ldots, n$, propose a regression that allows consistent estimation of σ_0^2 and σ_1^2. [*Hint*: You should first square the residuals.]

Computer Exercises

C1 Consider the following model to explain sleeping behavior:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u.$$

(i) Write down a model that allows the variance of u to differ between men and women. The variance should not depend on other factors.

(ii) Use the data in SLEEP75 to estimate the parameters of the model for heteroskedasticity. (You have to first estimate the *sleep* equation by OLS to obtain the OLS residuals.) Is the estimated variance of u higher for men or for women?

(iii) Is the variance of u statistically different for men and for women?

C2 (i) Use the data in HPRICE1 to obtain the heteroskedasticity-robust standard errors for equation (8.17). Discuss any important differences with the usual standard errors.

(ii) Repeat part (i) for equation (8.18).

(iii) What does this example suggest about heteroskedasticity and the transformation used for the dependent variable?

C3 Apply the full White test for heteroskedasticity [see equation (8.19)] to equation (8.18). Using the chi-square form of the statistic, obtain the p-value. What do you conclude?

C4 Use VOTE1 for this exercise.

(i) Estimate a model with *voteA* as the dependent variable and *prtystrA*, *democA*, log(*expendA*), and log(*expendB*) as independent variables. Obtain the OLS residuals, $\hat{u}_i$, and regress these on all of the independent variables. Explain why you obtain $R^2 = 0$.

(ii) Now, compute the Breusch-Pagan test for heteroskedasticity. Use the F statistic version and report the p-value.

(iii) Compute the special case of the White test for heteroskedasticity, again using the F statistic form. How strong is the evidence for heteroskedasticity now?

C5 Use the data in PNTSPRD for this exercise.

(i) The variable *sprdcvr* is a binary variable equal to one if the Las Vegas point spread for a college basketball game was covered. The expected value of *sprdcvr*, say μ, is the probability that the spread is covered in a randomly selected game. Test H_0: $\mu = .5$ against H_1: $\mu \neq .5$ at the 10%

significance level and discuss your findings. (*Hint:* This is easily done using a t test by regressing *sprdcvr* on an intercept only.)

(ii) How many games in the sample of 553 were played on a neutral court?

(iii) Estimate the linear probability model

$$sprdcvr = \beta_0 + \beta_1 favhome + \beta_2 neutral + \beta_3 fav25 + \beta_4 und25 + u$$

and report the results in the usual form. (Report the usual OLS standard errors and the heteroskedasticity-robust standard errors.) Which variable is most significant, both practically and statistically?

(iv) Explain why, under the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, there is no heteroskedasticity in the model.

(v) Use the usual F statistic to test the hypothesis in part (iv). What do you conclude?

(vi) Given the previous analysis, would you say that it is possible to systematically predict whether the Las Vegas spread will be covered using information available prior to the game?

C6 In Example 7.12, we estimated a linear probability model for whether a young man was arrested during 1986:

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u.$$

(i) Using the data in CRIME1, estimate this model by OLS and verify that all fitted values are strictly between zero and one. What are the smallest and largest fitted values?

(ii) Estimate the equation by weighted least squares, as discussed in Section 8-5.

(iii) Use the WLS estimates to determine whether *avgsen* and *tottime* are jointly significant at the 5% level.

C7 Use the data in LOANAPP for this exercise.

(i) Estimate the equation in part (iii) of Computer Exercise C8 in Chapter 7, computing the heteroskedasticity-robust standard errors. Compare the 95% confidence interval on β_{white} with the nonrobust confidence interval.

(ii) Obtain the fitted values from the regression in part (i). Are any of them less than zero? Are any of them greater than one? What does this mean about applying weighted least squares?

C8 Use the data set GPA1 for this exercise.

(i) Use OLS to estimate a model relating *colGPA* to *hsGPA*, *ACT*, *skipped*, and *PC*. Obtain the OLS residuals.

(ii) Compute the special case of the White test for heteroskedasticity. In the regression of $\hat{u}_i^2$ on $\widehat{colGPA}_i$, $\widehat{colGPA}_i^2$, obtain the fitted values, say $\hat{h}_i$.

(iii) Verify that the fitted values from part (ii) are all strictly positive. Then, obtain the weighted least squares estimates using weights $1/\hat{h}_i$. Compare the weighted least squares estimates for the effect of skipping lectures and the effect of PC ownership with the corresponding OLS estimates. What about their statistical significance?

(iv) In the WLS estimation from part (iii), obtain heteroskedasticity-robust standard errors. In other words, allow for the fact that the variance function estimated in part (ii) might be misspecified. (See Question 8.4.) Do the standard errors change much from part (iii)?

C9 In Example 8.7, we computed the OLS and a set of WLS estimates in a cigarette demand equation.

(i) Obtain the OLS estimates in equation (8.35).

(ii) Obtain the $\hat{h}_i$ used in the WLS estimation of equation (8.36) and reproduce equation (8.36). From this equation, obtain the *unweighted* residuals and fitted values; call these $\hat{u}_i$ and $\hat{y}_i$, respectively. (For example, in Stata®, the unweighted residuals and fitted values are given by default.)

(iii) Let $\check{u}_i = \hat{u}_i/\sqrt{\hat{h}_i}$ and $\check{y}_i = \hat{y}_i/\sqrt{\hat{h}_i}$ be the weighted quantities. Carry out the special case of the White test for heteroskedasticity by regressing $\check{u}_i^2$ on $\check{y}_i$, $\check{y}_i^2$, being sure to include an intercept, as always. Do you find heteroskedasticity in the weighted residuals?

(iv) What does the finding from part (iii) imply about the proposed form of heteroskedasticity used in obtaining (8.36)?

(v) Obtain valid standard errors for the WLS estimates that allow the variance function to be misspecified.

C10 Use the data set 401KSUBS for this exercise.

(i) Using OLS, estimate a linear probability model for *e401k*, using as explanatory variables *inc*, inc^2, *age*, age^2, and *male*. Obtain both the usual OLS standard errors and the heteroskedasticity-robust versions. Are there any important differences?

(ii) In the special case of the White test for heteroskedasticity, where we regress the squared OLS residuals on a quadratic in the OLS fitted values, $\hat{u}_i^2$ on $\hat{y}_i$, $\hat{y}_i^2$, $i = 1, \ldots, n$, argue that the probability limit of the coefficient on $\hat{y}_i$ should be one, the probability limit of the coefficient on $\hat{y}_i^2$ should be -1, and the probability limit of the intercept should be zero. {*Hint:* Remember that $\text{Var}(y|x_1, \ldots, x_k) = p(\mathbf{x})[1 - p(\mathbf{x})]$, where $p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$.}

(iii) For the model estimated from part (i), obtain the White test and see if the coefficient estimates roughly correspond to the theoretical values described in part (ii).

(iv) After verifying that the fitted values from part (i) are all between zero and one, obtain the weighted least squares estimates of the linear probability model. Do they differ in important ways from the OLS estimates?

C11 Use the data in 401KSUBS for this question, restricting the sample to $fsize = 1$.

(i) To the model estimated in Table 8.1, add the interaction term, $e401k \cdot inc$. Estimate the equation by OLS and obtain the usual and robust standard errors. What do you conclude about the statistical significance of the interaction term?

(ii) Now estimate the more general model by WLS using the same weights, $1/inc_i$, as in Table 8.1. Compute the usual and robust standard error for the WLS estimator. Is the interaction term statistically significant using the robust standard error?

(iii) Discuss the WLS coefficient on *e401k* in the more general model. Is it of much interest by itself? Explain.

(iv) Reestimate the model by WLS but use the interaction term $e401k \cdot (inc - 30)$; the average income in the sample is about 29.44. Now interpret the coefficient on *e401k*.

C12 Use the data in MEAP00 to answer this question.

(i) Estimate the model

$$math4 = \beta_0 + \beta_1 lunch + \beta_2 \log(enroll) + \beta_3 \log(exppp) + u$$

by OLS and obtain the usual standard errors and the fully robust standard errors. How do they generally compare?

(ii) Apply the special case of the White test for heteroskedasticity. What is the value of the F test? What do you conclude?

(iii) Obtain $\hat{g}_i$ as the fitted values from the regression $\log(\hat{u}_i^2)$ on $\widehat{math4}_i$, $\widehat{math4}_i^2$, where $\widehat{math4}_i$ are the OLS fitted values and the $\hat{u}_i$ are the OLS residuals. Let $\hat{h}_i = \exp(\hat{g}_i)$. Use the $\hat{h}_i$ to obtain WLS estimates. Are there big differences with the OLS coefficients?

(iv) Obtain the standard errors for WLS that allow misspecification of the variance function. Do these differ much from the usual WLS standard errors?

(v) For estimating the effect of spending on *math*4, does OLS or WLS appear to be more precise?

C13 Use the data in FERTIL2 to answer this question.

(i) Estimate the model

$$children = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educ + \beta_4 electric + \beta_5 urban + u$$

and report the usual and heteroskedasticity-robust standard errors. Are the robust standard errors always bigger than the nonrobust ones?

(ii) Add the three religious dummy variables and test whether they are jointly significant. What are the p-values for the nonrobust and robust tests?

(iii) From the regression in part (ii), obtain the fitted values $\hat{y}$ and the residuals, $\hat{u}$. Regress $\hat{u}^2$ on $\hat{y}$, $\hat{y}_2$ and test the joint significance of the two regressors. Conclude that heteroskedasticity is present in the equation for *children*.

(iv) Would you say the heteroskedasticity you found in part (iii) is practically important?

C14 Use the data in BEAUTY for this question.

(i) Using the data pooled for men and women, estimate the equation

$$lwage = \beta_0 + \beta_1 belavg + \beta_2 abvavg + \beta_3 female + \beta_4 educ + \beta_5 exper + \beta_5 exper^2 + u,$$

and report the results using heteroskedasticity-robust standard errors below coefficients. Are any of the coefficients surprising in either their signs or magnitudes? Is the coefficient on *female* practically large and statistically significant?

(ii) Add interactions of *female* with all other explanatory variables in the equation from part (i) (five interactions in all). Compute the usual F test of joint significance of the five interactions and a heteroskedasticity-robust version. Does using the heteroskedasticity-robust version change the outcome in any important way?

(iii) In the full model with interactions, determine whether those involving the looks variables—*female* • *belavg* and *female* • *abvavg*—are jointly significant. Are their coefficients practically small?

CHAPTER 9

More on Specification and Data Issues

In Chapter 8, we dealt with one failure of the Gauss-Markov assumptions. While heteroskedasticity in the errors can be viewed as a problem with a model, it is a relatively minor one. The presence of heteroskedasticity does not cause bias or inconsistency in the OLS estimators. Also, it is fairly easy to adjust confidence intervals and t and F statistics to obtain valid inference after OLS estimation, or even to get more efficient estimators by using weighted least squares.

In this chapter, we return to the much more serious problem of correlation between the error, u, and one or more of the explanatory variables. Remember from Chapter 3 that if u is, for whatever reason, correlated with the explanatory variable x_j, then we say that x_j is an **endogenous explanatory variable**. We also provide a more detailed discussion on three reasons why an explanatory variable can be endogenous; in some cases, we discuss possible remedies.

We have already seen in Chapters 3 and 5 that omitting a key variable can cause correlation between the error and some of the explanatory variables, which generally leads to bias and inconsistency in *all* of the OLS estimators. In the special case that the omitted variable is a function of an explanatory variable in the model, the model suffers from **functional form misspecification**.

We begin in the first section by discussing the consequences of functional form misspecification and how to test for it. In Section 9-2, we show how the use of proxy variables can solve, or at least mitigate, omitted variables bias. In Section 9-3, we derive and explain the bias in OLS that can arise under certain forms of **measurement error**. Additional data problems are discussed in Section 9-4.

All of the procedures in this chapter are based on OLS estimation. As we will see, certain problems that cause correlation between the error and some explanatory variables cannot be solved by using OLS on a single cross section. We postpone a treatment of alternative estimation methods until Part 3.

9-1 Functional Form Misspecification

A multiple regression model suffers from functional form misspecification when it does not properly account for the relationship between the dependent and the observed explanatory variables. For example, if hourly wage is determined by $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$, but we omit the squared experience term, $exper^2$, then we are committing a functional form misspecification. We already know from Chapter 3 that this generally leads to biased estimators of β_0, β_1, and β_2. (We do not estimate β_3 because $exper^2$ is excluded from the model.) Thus, misspecifying how *exper* affects $\log(wage)$ generally results in a biased estimator of the return to education, β_1. The amount of this bias depends on the size of β_3 and the correlation among *educ*, *exper*, and $exper^2$.

Things are worse for estimating the return to experience: even if we could get an unbiased estimator of β_2, we would not be able to estimate the return to experience because it equals $\beta_2 + 2\beta_3 exper$ (in decimal form). Just using the biased estimator of β_2 can be misleading, especially at extreme values of *exper*.

As another example, suppose the $\log(wage)$ equation is

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \beta_5 female \cdot educ + u, \qquad [9.1]$$

where *female* is a binary variable. If we omit the interaction term, $female \cdot educ$, then we are misspecifying the functional form. In general, we will not get unbiased estimators of any of the other parameters, and because the return to education depends on gender, it is not clear what return we would be estimating by omitting the interaction term.

Omitting functions of independent variables is not the only way that a model can suffer from misspecified functional form. For example, if (9.1) is the true model satisfying the first four Gauss-Markov assumptions, but we use *wage* rather than $\log(wage)$ as the dependent variable, then we will not obtain unbiased or consistent estimators of the partial effects. The tests that follow have some ability to detect this kind of functional form problem, but there are better tests that we will mention in the subsection on testing against nonnested alternatives.

Misspecifying the functional form of a model can certainly have serious consequences. Nevertheless, in one important respect, the problem is minor: by definition, we have data on all the necessary variables for obtaining a functional relationship that fits the data well. This can be contrasted with the problem addressed in the next section, where a key variable is omitted on which we cannot collect data.

We already have a very powerful tool for detecting misspecified functional form: the F test for joint exclusion restrictions. It often makes sense to add quadratic terms of any significant variables to a model and to perform a joint test of significance. If the additional quadratics are significant, they can be added to the model (at the cost of complicating the interpretation of the model). However, significant quadratic terms can be symptomatic of other functional form problems, such as using the level of a variable when the logarithm is more appropriate, or vice versa. It can be difficult to pinpoint the precise reason that a functional form is misspecified. Fortunately, in many cases, using logarithms of certain variables and adding quadratics are sufficient for detecting many important nonlinear relationships in economics.

EXAMPLE 9.1 Economic Model of Crime

Table 9.1 contains OLS estimates of the economic model of crime (see Example 8.3). We first estimate the model without any quadratic terms; those results are in column (1).

In column (2), the squares of *pcnv*, *ptime86*, and *inc86* are added; we chose to include the squares of these variables because each level term is significant in column (1). The variable *qemp86* is a discrete variable taking on only five values, so we do not include its square in column (2).

TABLE 9.1 Dependent Variable: *narr86*

Independent Variables	(1)	(2)
pcnv	−.133 (.040)	.553 (.154)
$pcnv^2$	—	−.730 (.156)
avgsen	−.011 (.012)	−.017 (.012)
tottime	.012 (.009)	.012 (.009)
ptime86	−.041 (.009)	.287 (.004)
$ptime86^2$	—	−.0296 (.0039)
qemp86	−.051 (.014)	−.014 (.017)
inc86	−.0015 (.0003)	−.0034 (.0008)
$inc86^2$	—	−.000007 (.000003)
black	.327 (.045)	.292 (.045)
hispan	.194 (.040)	.164 (.039)
intercept	.569 (.036)	.505 (.037)
Observations	2,725	2,725
R-squared	.0723	.1035

Each of the squared terms is significant, and together they are jointly very significant ($F = 31.37$, with $df = 3$ and 2,713; the p-value is essentially zero). Thus, it appears that the initial model overlooked some potentially important nonlinearities.

GOING FURTHER 9.1

Why do we not include the squares of *black* and *hispan* in column (2) of Table 9.1?

Would it make sense to add interactions of *black* and *hispan* with some of the other variables reported in the table?

The presence of the quadratics makes interpreting the model somewhat difficult. For example, *pcnv* no longer has a strict deterrent effect: the relationship between *narr86* and *pcnv* is positive up until *pcnv* = .365, and then the relationship is negative. We might conclude that there is little or no deterrent effect at lower values of *pcnv*; the effect only kicks in at higher prior conviction rates. We would have to use more sophisticated functional forms than the quadratic to verify this conclusion. It may be that *pcnv* is not entirely exogenous. For example, men who have not been convicted in the past (so that *pcnv* = 0) are perhaps casual criminals, and so they are less likely to be arrested in 1986. This could be biasing the estimates.

Similarly, the relationship between *narr86* and *ptime86* is positive up until *ptime86* = 4.85 (almost five months in prison), and then the relationship is negative. The vast majority of men in the sample spent no time in prison in 1986, so again we must be careful in interpreting the results.

Legal income has a negative effect on *narr86* until *inc86* = 242.85; because income is measured in hundreds of dollars, this means an annual income of $24,285. Only 46 of the men in the sample have incomes above this level. Thus, we can conclude that *narr86* and *inc86* are negatively related with a diminishing effect.

Example 9.1 is a tricky functional form problem due to the nature of the dependent variable. Other models are theoretically better suited for handling dependent variables taking on a small number of integer values. We will briefly cover these models in Chapter 17.

9-1a RESET as a General Test for Functional Form Misspecification

Some tests have been proposed to detect general functional form misspecification. Ramsey's (1969) **regression specification error test (RESET)** has proven to be useful in this regard.

The idea behind RESET is fairly simple. If the original model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \qquad [9.2]$$

satisfies MLR.4, then no nonlinear functions of the independent variables should be significant when added to equation (9.2). In Example 9.1, we added quadratics in the significant explanatory variables. Although this often detects functional form problems, it has the drawback of using up many degrees of freedom if there are many explanatory variables in the original model (much as the straight form of the White test for heteroskedasticity consumes degrees of freedom). Further, certain kinds of neglected nonlinearities will not be picked up by adding quadratic terms. RESET adds polynomials in the OLS fitted values to equation (9.2) to detect general kinds of functional form misspecification.

To implement RESET, we must decide how many functions of the fitted values to include in an expanded regression. There is no right answer to this question, but the squared and cubed terms have proven to be useful in most applications.

Let $\hat{y}$ denote the OLS fitted values from estimating (9.2). Consider the expanded equation

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error. \qquad [9.3]$$

This equation seems a little odd, because functions of the fitted values from the initial estimation now appear as explanatory variables. In fact, we will not be interested in the estimated parameters from (9.3); we only use this equation to test whether (9.2) has missed important nonlinearities. The thing to remember is that $\hat{y}^2$ and $\hat{y}^3$ are just nonlinear functions of the x_j.

The null hypothesis is that (9.2) is correctly specified. Thus, RESET is the F statistic for testing H_0: $\delta_1 = 0, \delta_2 = 0$ in the expanded model (9.3). A significant F statistic suggests some sort of functional form problem. The distribution of the F statistic is approximately $F_{2,n-k-3}$ in large samples under the null hypothesis (and the Gauss-Markov assumptions). The *df* in the expanded equation (9.3) is $n - k - 1 - 2 = n - k - 3$. An *LM* version is also available (and the chi-square distribution will have two *df*). Further, the test can be made robust to heteroskedasticity using the methods discussed in Section 8-2.

EXAMPLE 9.2 Housing Price Equation

We estimate two models for housing prices. The first one has all variables in level form:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u. \qquad [9.4]$$

The second one uses the logarithms of all variables except *bdrms*:

$$\log(price) = \beta_0 + \beta_1 llotsize + \beta_2 lsqrft + \beta_3 bdrms + u. \qquad [9.5]$$

Using $n = 88$ houses in HPRICE1, the RESET statistic for equation (9.4) turns out to be 4.67; this is the value of an $F_{2,82}$ random variable $(n = 88, k = 3)$, and the associated p-value is .012. This is evidence of functional form misspecification in (9.4).

The RESET statistic in (9.5) is 2.56, with p-value = .084. Thus, we do not reject (9.5) at the 5% significance level (although we would at the 10% level). On the basis of RESET, the log-log model in (9.5) is preferred.

In the previous example, we tried two models for explaining housing prices. One was rejected by RESET, while the other was not (at least at the 5% level). Often, things are not so simple. A drawback with RESET is that it provides no real direction on how to proceed if the model is rejected. Rejecting (9.4) by using RESET does not immediately suggest that (9.5) is the next step. Equation (9.5) was estimated because constant elasticity models are easy to interpret and can have nice statistical properties. In this example, it so happens that it passes the functional form test as well.

Some have argued that RESET is a very general test for model misspecification, including unobserved omitted variables and heteroskedasticity. Unfortunately, such use of RESET is largely misguided. It can be shown that RESET has no power for detecting omitted variables whenever they have expectations that are linear in the included independent variables in the model [see Wooldridge (2001, Section 2-1) for a precise statement]. Further, if the functional form is properly specified, RESET has no power for detecting heteroskedasticity. The bottom line is that RESET is a functional form test, and nothing more.

9-1b Tests against Nonnested Alternatives

Obtaining tests for other kinds of functional form misspecification—for example, trying to decide whether an independent variable should appear in level or logarithmic form—takes us outside the realm of classical hypothesis testing. It is possible to test the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad [9.6]$$

against the model

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u, \qquad [9.7]$$

and vice versa. However, these are **nonnested models** (see Chapter 6), and so we cannot simply use a standard F test. Two different approaches have been suggested. The first is to construct a comprehensive model that contains each model as a special case and then to test the restrictions that led to each of the models. In the current example, the comprehensive model is

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log(x_1) + \gamma_4 \log(x_2) + u. \qquad [9.8]$$

We can first test H_0: $\gamma_3 = 0, \gamma_4 = 0$ as a test of (9.6). We can also test H_0: $\gamma_1 = 0, \gamma_2 = 0$ as a test of (9.7). This approach was suggested by Mizon and Richard (1986).

Another approach has been suggested by Davidson and MacKinnon (1981). They point out that if model (9.6) holds with $E(u|x_1, x_2) = 0$, the fitted values from the other model, (9.7), should be insignificant when added to equation (9.6). Therefore, to test whether (9.6) is the correct model, we first estimate model (9.7) by OLS to obtain the fitted values; call these $\check{y}$. The **Davidson-MacKinnon test** is obtained from the t statistic on $\check{y}$ in the auxiliary equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \check{y} + error.$$

Because the $\check{y}$ are just nonlinear functions of x_1 and x_2, they should be insignificant if (9.6) is the correct conditional mean model. Therefore, a significant t statistic (against a two-sided alternative) is a rejection of (9.6).

Similarly, if $\hat{y}$ denotes the fitted values from estimating (9.6), the test of (9.7) is the t statistic on $\hat{y}$ in the model

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_1 \hat{y} + error;$$

a significant t statistic is evidence against (9.7). The same two tests can be used for testing any two nonnested models with the same dependent variable.

There are a few problems with nonnested testing. First, a clear winner need not emerge. Both models could be rejected or neither model could be rejected. In the latter case, we can use the adjusted R-squared to choose between them. If both models are rejected, more work needs to be done. However, it is important to know the practical consequences from using one form or the other: if the effects of key independent variables on y are not very different, then it does not really matter which model is used.

A second problem is that rejecting (9.6) using, say, the Davidson-MacKinnon test does not mean that (9.7) is the correct model. Model (9.6) can be rejected for a variety of functional form misspecifications.

An even more difficult problem is obtaining nonnested tests when the competing models have different dependent variables. The leading case is y versus $\log(y)$. We saw in Chapter 6 that just obtaining goodness-of-fit measures that can be compared requires some care. Tests have been proposed to solve this problem, but they are beyond the scope of this text. [See Wooldridge (1994a) for a test that has a simple interpretation and is easy to implement.]

9-2 Using Proxy Variables for Unobserved Explanatory Variables

A more difficult problem arises when a model excludes a key variable, usually because of data unavailability. Consider a wage equation that explicitly recognizes that ability (*abil*) affects $\log(wage)$:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u. \qquad [9.9]$$

This model shows explicitly that we want to hold ability fixed when measuring the return to *educ* and *exper*. If, say, *educ* is correlated with *abil*, then putting *abil* in the error term causes the OLS estimator of β_1 (and β_2) to be biased, a theme that has appeared repeatedly.

Our primary interest in equation (9.9) is in the slope parameters β_1 and β_2. We do not really care whether we get an unbiased or consistent estimator of the intercept β_0; as we will see shortly, this is not usually possible. Also, we can never hope to estimate β_3 because *abil* is not observed; in fact, we would not know how to interpret β_3 anyway, because ability is at best a vague concept.

How can we solve, or at least mitigate, the omitted variables bias in an equation like (9.9)? One possibility is to obtain a **proxy variable** for the omitted variable. Loosely speaking, a proxy variable is something that is related to the unobserved variable that we would like to control for in our analysis. In the wage equation, one possibility is to use the intelligence quotient, or IQ, as a proxy for ability. This *does not* require IQ to be the same thing as ability; what we need is for IQ to be correlated with ability, something we clarify in the following discussion.

All of the key ideas can be illustrated in a model with three independent variables, two of which are observed:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u. \qquad [9.10]$$

We assume that data are available on y, x_1, and x_2—in the wage example, these are $\log(wage)$, *educ*, and *exper*, respectively. The explanatory variable x_3^* is unobserved, but we have a proxy variable for x_3^*. Call the proxy variable x_3.

What do we require of x_3? At a minimum, it should have some relationship to x_3^*. This is captured by the simple regression equation

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3, \qquad [9.11]$$

where v_3 is an error due to the fact that x_3^* and x_3 are not exactly related. The parameter δ_3 measures the relationship between x_3^* and x_3; typically, we think of x_3^* and x_3 as being positively related, so that $\delta_3 > 0$. If $\delta_3 = 0$, then x_3 is not a suitable proxy for x_3^*. The intercept δ_0 in (9.11), which can be positive or negative, simply allows x_3^* and x_3 to be measured on different scales. (For example, unobserved ability is certainly not required to have the same average value as IQ in the U.S. population.)

How can we use x_3 to get unbiased (or at least consistent) estimators of β_1 and β_2? The proposal is to pretend that x_3 and x_3^* are the same, so that we run the regression of

$$y \text{ on } x_1, x_2, x_3. \qquad [9.12]$$

We call this the **plug-in solution to the omitted variables problem** because x_3 is just plugged in for x_3^* before we run OLS. If x_3 is truly related to x_3^*, this seems like a sensible thing. However, because x_3 and x_3^* are not the same, we should determine when this procedure does in fact give consistent estimators of β_1 and β_2.

The assumptions needed for the plug-in solution to provide consistent estimators of β_1 and β_2 can be broken down into assumptions about u and v_3:

(1) The error u is uncorrelated with x_1, x_2, and x_3^*, which is just the standard assumption in model (9.10). In addition, u is uncorrelated with x_3. This latter assumption just means that x_3 is irrelevant in the population model, once x_1, x_2, and x_3^* have been included. This is essentially true by definition, as x_3 is a proxy variable for x_3^*: it is x_3^* that directly affects y, not x_3. Thus, the assumption that u is uncorrelated with x_1, x_2, x_3^*, and x_3 is not very controversial. (Another way to state this assumption is that the expected value of u, given all these variables, is zero.)

(2) The error v_3 is uncorrelated with x_1, x_2, and x_3. Assuming that v_3 is uncorrelated with x_1 and x_2 requires x_3 to be a "good" proxy for x_3^*. This is easiest to see by writing the analog of these assumptions in terms of conditional expectations:

$$\mathrm{E}(x_3^*|x_1,x_2,x_3) = \mathrm{E}(x_3^*|x_3) = \delta_0 + \delta_3 x_3. \qquad [9.13]$$

The first equality, which is the most important one, says that, once x_3 is controlled for, the expected value of x_3^* does not depend on x_1 or x_2. Alternatively, x_3^* has zero correlation with x_1 and x_2 once x_3 is partialled out.

In the wage equation (9.9), where IQ is the proxy for ability, condition (9.13) becomes

$$\mathrm{E}(abil|educ,exper,IQ) = \mathrm{E}(abil|IQ) = \delta_0 + \delta_3 IQ.$$

Thus, the average level of ability only changes with IQ, not with *educ* and *exper*. Is this reasonable? Maybe it is not exactly true, but it may be close to being true. It is certainly worth including IQ in the wage equation to see what happens to the estimated return to education.

We can easily see why the previous assumptions are enough for the plug-in solution to work. If we plug equation (9.11) into equation (9.10) and do simple algebra, we get

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3.$$

Call the composite error in this equation $e = u + \beta_3 v_3$; it depends on the error in the model of interest, (9.10), and the error in the proxy variable equation, v_3. Because u and v_3 both have zero mean and each is uncorrelated with x_1, x_2, and x_3, e also has zero mean and is uncorrelated with x_1, x_2, and x_3. Write this equation as

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e,$$

where $\alpha_0 = (\beta_0 + \beta_3 \delta_0)$ is the new intercept and $\alpha_3 = \beta_3 \delta_3$ is the slope parameter on the proxy variable x_3. As we alluded to earlier, when we run the regression in (9.12), we will not get unbiased estimators of β_0 and β_3; instead, we will get unbiased (or at least consistent) estimators of α_0, β_1, β_2, and α_3. The important thing is that we get good estimates of the parameters β_1 and β_2.

In most cases, the estimate of α_3 is actually more interesting than an estimate of β_3 anyway. For example, in the wage equation, α_3 measures the return to wage given one more point on IQ score.

EXAMPLE 9.3 IQ as a Proxy for Ability

The file WAGE2, from Blackburn and Neumark (1992), contains information on monthly earnings, education, several demographic variables, and IQ scores for 935 men in 1980. As a method to account for omitted ability bias, we add *IQ* to a standard log wage equation. The results are shown in Table 9.2.

Our primary interest is in what happens to the estimated return to education. Column (1) contains the estimates without using *IQ* as a proxy variable. The estimated return to education is 6.5%. If we think omitted ability is positively correlated with *educ*, then we assume that this estimate is too high. (More precisely, the average estimate across all random samples would be too high.) When *IQ* is

TABLE 9.2 Dependent Variable: log(*wage*)

Independent Variables	(1)	(2)	(3)
educ	.065 (.006)	.054 (.007)	.018 (.041)
exper	.014 (.003)	.014 (.003)	.014 (.003)
tenure	.012 (.002)	.011 (.002)	.011 (.002)
married	.199 (.039)	.200 (.039)	.201 (.039)
south	−.091 (.026)	−.080 (.026)	−.080 (.026)
urban	.184 (.027)	.182 (.027)	.184 (.027)
black	−.188 (.038)	−.143 (.039)	−.147 (.040)
IQ	—	.0036 (.0010)	−.0009 (.0052)
educ · *IQ*	—	—	.00034 (.00038)
intercept	5.395 (.113)	5.176 (.128)	5.648 (.546)
Observations	935	935	935
R-squared	.253	.263	.263

added to the equation, the return to education falls to 5.4%, which corresponds with our prior beliefs about omitted ability bias.

The effect of IQ on socioeconomic outcomes has been documented in the controversial book *The Bell Curve*, by Herrnstein and Murray (1994). Column (2) shows that IQ does have a statistically significant, positive effect on earnings, after controlling for several other factors. Everything else being equal, an increase of 10 IQ points is predicted to raise monthly earnings by 3.6%. The standard deviation of IQ in the U.S. population is 15, so a one standard deviation increase in IQ is associated with higher earnings of 5.4%. This is identical to the predicted increase in wage due to another year of education. It is clear from column (2) that education still has an important role in increasing earnings, even though the effect is not as large as originally estimated.

Some other interesting observations emerge from columns (1) and (2). Adding *IQ* to the equation only increases the *R*-squared from .253 to .263. Most of the variation in $\log(wage)$ is not explained by the factors in column (2). Also, adding *IQ* to the equation does not eliminate the estimated earnings difference between black and white men: a black man with the same IQ, education, experience, and so on, as a white man is predicted to earn about 14.3% less, and the difference is very statistically significant.

GOING FURTHER 9.2

What do you make of the small and statistically insignificant coefficient on *educ* in column (3) of Table 9.2? (*Hint:* When $educ \cdot IQ$ is in the equation, what is the interpretation of the coefficient on *educ*?)

Column (3) in Table 9.2 includes the interaction term $educ \cdot IQ$. This allows for the possibility that *educ* and *abil* interact in determining $\log(wage)$. We might think that the return to education is higher for people with more ability, but this turns out not to be the case: the interaction term is not significant, and its addition makes *educ* and *IQ* individually insignificant while complicating the model. Therefore, the estimates in column (2) are preferred.

There is no reason to stop at a single proxy variable for ability in this example. The data set WAGE2 also contains a score for each man on the *Knowledge of the World of Work* (KWW) test. This provides a different measure of ability, which can be used in place of IQ or along with IQ, to estimate the return to education (see Computer Exercise C2).

It is easy to see how using a proxy variable can still lead to bias if the proxy variable does not satisfy the preceding assumptions. Suppose that, instead of (9.11), the unobserved variable, x_3^*, is related to all of the observed variables by

$$x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3, \qquad [9.14]$$

where v_3 has a zero mean and is uncorrelated with x_1, x_2, and x_3. Equation (9.11) assumes that δ_1 and δ_2 are both zero. By plugging equation (9.14) into (9.10), we get

$$\begin{aligned} y &= (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 \\ &\quad + \beta_3\delta_3 x_3 + u + \beta_3 v_3, \end{aligned} \qquad [9.15]$$

from which it follows that $\text{plim}(\hat{\beta}_1) = \beta_1 + \beta_3\delta_1$ and $\text{plim}(\hat{\beta}_2) = \beta_2 + \beta_3\delta_2$. [This follows because the error in (9.15), $u + \beta_3 v_3$, has zero mean and is uncorrelated with x_1, x_2, and x_3.] In the previous example where $x_1 = educ$ and $x_3^* = abil$, $\beta_3 > 0$, so there is a positive bias (inconsistency) if *abil* has a positive partial correlation with *educ* ($\delta_1 > 0$). Thus, we could still be getting an upward bias in the return to education by using *IQ* as a proxy for *abil* if *IQ* is not a good proxy. But we can reasonably hope that this bias is smaller than if we ignored the problem of omitted ability entirely.

A complaint that is sometimes aired about including variables such as *IQ* in a regression that includes *educ* is that it exacerbates the problem of multicollinearity, likely leading to a less precise estimate of β_{educ}. But this complaint misses two important points. First, the inclusion of *IQ* reduces the error variance because the part of ability explained by *IQ* has been removed from the error. Typically,

this will be reflected in a smaller standard error of the regression (although it need not get smaller because of its degrees-of-freedom adjustment). Second, and most importantly, the added multicollinearity is a necessary evil if we want to get an estimator of β_{educ} with less bias: the reason *educ* and *IQ* are correlated is that *educ* and *abil* are thought to be correlated, and *IQ* is a proxy for *abil*. If we could observe *abil* we would include it in the regression, and of course, there would be unavoidable multicollinearity caused by correlation between *educ* and *abil*.

Proxy variables can come in the form of binary information as well. In Example 7.9 [see equation (7.15)], we discussed Krueger's (1993) estimates of the return to using a computer on the job. Krueger also included a binary variable indicating whether the worker uses a computer at home (as well as an interaction term between computer usage at work and at home). His primary reason for including computer usage at home in the equation was to proxy for unobserved "technical ability" that could affect wage directly and be related to computer usage at work.

9-2a Using Lagged Dependent Variables as Proxy Variables

In some applications, like the earlier wage example, we have at least a vague idea about which unobserved factor we would like to control for. This facilitates choosing proxy variables. In other applications, we suspect that one or more of the independent variables is correlated with an omitted variable, but we have no idea how to obtain a proxy for that omitted variable. In such cases, we can include, as a control, the value of the dependent variable from an earlier time period. This is especially useful for policy analysis.

Using a **lagged dependent variable** in a cross-sectional equation increases the data requirements, but it also provides a simple way to account for historical factors that cause *current* differences in the dependent variable that are difficult to account for in other ways. For example, some cities have had high crime rates in the past. Many of the same unobserved factors contribute to both high current and past crime rates. Likewise, some universities are traditionally better in academics than other universities. Inertial effects are also captured by putting in lags of *y*.

Consider a simple equation to explain city crime rates:

$$crime = \beta_0 + \beta_1 unem + \beta_2 expend + \beta_3 crime_{-1} + u, \qquad [9.16]$$

where *crime* is a measure of per capita crime, *unem* is the city unemployment rate, *expend* is per capita spending on law enforcement, and $crime_{-1}$ indicates the crime rate measured in some earlier year (this could be the past year or several years ago). We are interested in the effects of *unem* on *crime*, as well as of law enforcement expenditures on crime.

What is the purpose of including $crime_{-1}$ in the equation? Certainly, we expect that $\beta_3 > 0$ because crime has inertia. But the main reason for putting this in the equation is that cities with high historical crime rates may spend more on crime prevention. Thus, factors unobserved by us (the econometricians) that affect *crime* are likely to be correlated with *expend* (and *unem*). If we use a pure cross-sectional analysis, we are unlikely to get an unbiased estimator of the causal effect of law enforcement expenditures on crime. But, by including $crime_{-1}$ in the equation, we can at least do the following experiment: if two cities have the same previous crime rate and current unemployment rate, then β_2 measures the effect of another dollar of law enforcement on crime.

EXAMPLE 9.4 City Crime Rates

We estimate a constant elasticity version of the crime model in equation (9.16) (*unem*, because it is a percentage, is left in level form). The data in CRIME2 are from 46 cities for the year 1987. The crime rate is also available for 1982, and we use that as an additional independent variable in trying to control for city unobservables that affect crime and may be correlated with current law enforcement expenditures. Table 9.3 contains the results.

TABLE 9.3 Dependent Variable: $\log(crmrte_{87})$

Independent Variables	(1)	(2)
$unem_{87}$	−.029 (.032)	.009 (.020)
$\log(lawexpc_{87})$	.203 (.173)	−.140 (.109)
$\log(crmrte_{82})$	—	1.194 (.132)
intercept	3.34 (1.25)	.076 (.821)
Observations	46	46
R-squared	.057	.680

Without the lagged crime rate in the equation, the effects of the unemployment rate and expenditures on law enforcement are counterintuitive; neither is statistically significant, although the t statistic on $\log(lawexpc_{87})$ is 1.17. One possibility is that increased law enforcement expenditures improve reporting conventions, and so more crimes are *reported*. But it is also likely that cities with high recent crime rates spend more on law enforcement.

Adding the log of the crime rate from five years earlier has a large effect on the expenditures coefficient. The elasticity of the crime rate with respect to expenditures becomes −.14, with $t = -1.28$. This is not strongly significant, but it suggests that a more sophisticated model with more cities in the sample could produce significant results.

Not surprisingly, the current crime rate is strongly related to the past crime rate. The estimate indicates that if the crime rate in 1982 was 1% higher, then the crime rate in 1987 is predicted to be about 1.19% higher. We cannot reject the hypothesis that the elasticity of current crime with respect to past crime is unity $[t = (1.194 - 1)/.132 \approx 1.47]$. Adding the past crime rate increases the explanatory power of the regression markedly, but this is no surprise. The primary reason for including the lagged crime rate is to obtain a better estimate of the ceteris paribus effect of $\log(lawexpc_{87})$ on $\log(crmrte_{87})$.

The practice of putting in a lagged y as a general way of controlling for unobserved variables is hardly perfect. But it can aid in getting a better estimate of the effects of policy variables on various outcomes. When the data are available, additional lags also can be included.

Adding lagged values of y is not the only way to use two years of data to control for omitted factors. When we discuss panel data methods in Chapters 13 and 14, we will cover other ways to use repeated data on the same cross-sectional units at different points in time.

9-2b A Different Slant on Multiple Regression

The discussion of proxy variables in this section suggests an alternative way of interpreting a multiple regression analysis when we do not necessarily observe all relevant explanatory variables. Until now, we have specified the population model of interest with an additive error, as in equation (9.9). Our discussion of that example hinged upon whether we have a suitable proxy variable (IQ score in this case, other test scores more generally) for the unobserved explanatory variable, which we called "ability."

A less structured, more general approach to multiple regression is to forego specifying models with unobservables. Rather, we begin with the premise that we have access to a set of observable explanatory variables—which includes the variable of primary interest, such as years of schooling,

and controls, such as observable test scores. We then model the mean of y conditional on the observed explanatory variables. For example, in the wage example with *lwage* denoting $\log(wage)$, we can estimate $\mathrm{E}(lwage|educ, exper, tenure, south, urban, black, IQ)$—exactly what is reported in Table 9.2. The difference now is that we set our goals more modestly. Namely, rather than introduce the nebulous concept of "ability" in equation (9.9), we state from the outset that we will estimate the ceteris paribus effect of education holding *IQ* (and the other observed factors) fixed. There is no need to discuss whether *IQ* is a suitable proxy for ability. Consequently, while we may not be answering the question underlying equation (9.9), we are answering a question of interest: if two people have the same *IQ* levels (and same values of experience, tenure, and so on), yet they differ in education levels by a year, what is the expected difference in their log wages?

As another example, if we include as an explanatory variable the poverty rate in a school-level regression to assess the effects of spending on standardized test scores, we should recognize that the poverty rate only crudely captures the relevant differences in children and parents across schools. But often it is all we have, and it is better to control for the poverty rate than to do nothing because we cannot find suitable proxies for student "ability," parental "involvement," and so on. Almost certainly controlling for the poverty rate gets us closer to the ceteris paribus effects of spending than if we leave the poverty rate out of the analysis.

In some applications of regression analysis, we are interested simply in predicting the outcome, y, given a set of explanatory variables, $(x_1, \ldots, x_k)$. In such cases, it makes little sense to think in terms of "bias" in estimated coefficients due to omitted variables. Instead, we should focus on obtaining a model that predicts as well as possible, and make sure we do not include as regressors variables that cannot be observed at the time of prediction. For example, an admissions officer for a college or university might be interested in predicting success in college, as measured by grade point average, in terms of variables that can be measured at application time. Those variables would include high school performance (maybe just grade point average, but perhaps performance in specific kinds of courses), standardized test scores, participation in various activities (such as debate or math club), and even family background variables. We would not include a variable measuring college class attendance because we do not observe attendance in college at application time. Nor would we wring our hands about potential "biases" caused by omitting an attendance variable: we have no interest in, say, measuring the effect of high school GPA holding attendance in college fixed. Likewise, we would not worry about biases in coefficients because we cannot observe factors such as motivation. Naturally, for predictive purposes it would probably help substantially if we had a measure of motivation, but in its absence we fit the best model we can with observed explanatory variables.

9-2c Potential Outcomes and Proxy Variables

The notion of proxy variables can be related to the potential outcomes framework that we introduced in Sections 2-7, 3-7, and 4-7, covered in more generality in Section 7-6. Let $y(0)$ and $y(1)$ denote the potential outcomes and w the binary treatment indicator. When we include explanatory variables $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ in a regression that includes w as an explanatory variable, one way to think about what we are doing is we are using $\mathbf{x}$ as a set of proxy variables for the unobserved factors that affect the potential outcomes, y(0) and y(1), and might also be related to the participation decision ($w = 1$ or $w = 0$). Write

$$y(0) = \mu_0 + v(0)$$
$$y(1) = \mu_1 + v(1)$$

where μ_0 and μ_1 are the two counterfactual means and $\tau_{ate} = \mu_1 - \mu_0$ is the average treatment effect. The problem of selection into participation means that w can be related to the unobservables $v(0)$ and $v(1)$. The ignorability or unconfoundedness assumption discussed in Sections 3-7 and 7-6 is that,

conditional on $\mathbf{x}$, w is independent of $[v(0), v(1)]$. This is essentially the assumption that the elements of $\mathbf{x}$ act as suitable proxies for the unobservables. Assuming linear functional forms as in Section 7-6,

$$\mathrm{E}[v(0)|w,\mathbf{x}] = \mathrm{E}[v(0)|\mathbf{x}] = (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\beta}_0 \text{ and}$$
$$\mathrm{E}[v(1)|w,\mathbf{x}] = \mathrm{E}[v(1)|\mathbf{x}] = (\mathbf{x} - \boldsymbol{\eta})\boldsymbol{\beta}_1,$$

where $\boldsymbol{\eta} = \mathrm{E}(\mathbf{x})$. The first equalities in both equations, which we recognize as implications of the conditional independence or unconfoundedness assumption, is effectively the same as the proxy variable condition: conditional on $\mathbf{x}$, w is unrelated to unobserved factors that affect $[y(0),y(1)]$. From Section 7-6 we know that unconfoundedness plus the linear functional form leads to a regression with interaction terms,

$$y_i \text{ on } w_i, \mathbf{x}_i, w_i \cdot (\mathbf{x} - \bar{\mathbf{x}}), i = 1, \ldots, n,$$

where $\bar{\mathbf{x}}$ is the vector of sample averages and the regression is across the entire sample. The coefficient on w_i is $\hat{\tau}_{ate}$, the estimate of the average treatment effect. See Section 7-6 for further discussion.

9-3 Models with Random Slopes

In our treatment of regression so far, we have assumed that the slope coefficients are the same across individuals in the population, or that, if the slopes differ, they differ by measurable characteristics, in which case we are led to regression models containing interaction terms. For example, as we saw in Section 7-4, we can allow the return to education to differ by men and women by interacting education with a gender dummy in a log wage equation.

Here we are interested in a related but different question: What if the partial effect of a variable depends on unobserved factors that vary by population unit? If we have only one explanatory variable, x, we can write a general model (for a random draw, i, from the population, for emphasis) as

$$y_i = a_i + b_i x_i, \qquad [9.17]$$

where a_i is the intercept for unit i and b_i is the slope. In the simple regression model from Chapter 2 we assumed $b_i = \beta$ and labeled a_i as the error, u_i. The model in (9.17) is sometimes called a **random coefficient model** or **random slope model** because the unobserved slope coefficient, b_i, is viewed as a random draw from the population along with the observed data, (x_i,y_i), and the unobserved intercept, a_i. As an example, if $y_i = \log(wage_i)$ and $x_i = educ_i$, then (9.17) allows the return to education, b_i, to vary by person. If, say, b_i contains unmeasured ability (just as a_i would), the partial effect of another year of schooling can depend on ability.

With a random sample of size n, we (implicitly) draw n values of b_i along with n values of a_i (and the observed data on x and y). Naturally, we cannot estimate a slope—or, for that matter, an intercept—for each i. But we can hope to estimate the average slope (and average intercept), where the average is across the population. Therefore, define $\alpha = \mathrm{E}(a_i)$ and $\beta = \mathrm{E}(b_i)$. Then β is the average of the partial effect of x on y, and so we call β the **average partial effect (APE)**, or the **average marginal effect (AME)**. In the context of a log wage equation, β is the average return to a year of schooling in the population.

If we write $a_i = \alpha + c_i$ and $b_i = \beta + d_i$, then d_i is the individual-specific deviation from the APE. By construction, $\mathrm{E}(c_i) = 0$ and $\mathrm{E}(d_i) = 0$. Substituting into (9.17) gives

$$y_i = \alpha + \beta x_i + c_i + d_i x_i = \alpha + \beta x_i + u_i, \qquad [9.18]$$

where $u_i = c_i + d_i x_i$. (To make the notation easier to follow, we now use α, the mean value of a_i, as the intercept, and β, the mean of b_i, as the slope.) In other words, we can write the random coefficient as a constant coefficient model but where the error term contains an interaction between an unobservable, d_i, and the observed explanatory variable, x_i.

When would a simple regression of y_i on x_i provide an unbiased estimate of β (and α)? We can apply the result for unbiasedness from Chapter 2. If $E(u_i|x_i) = 0$, then OLS is generally unbiased. When $u_i = c_i + d_i x_i$, sufficient is $E(c_i|x_i) = E(c_i) = 0$ and $E(d_i|x_i) = E(d_i) = 0$. We can write these in terms of the unit-specific intercept and slope as

$$E(a_i|x_i) = E(a_i) \quad \text{and} \quad E(b_i|x_i) = E(b_i); \tag{9.19}$$

that is, a_i and b_i are both mean independent of x_i. This is a useful finding: if we allow for unit-specific slopes, OLS consistently estimates the population average of those slopes when they are mean independent of the explanatory variable. (See Problem 6 for a weaker set of conditions that imply consistency of OLS.)

The error term in (9.18) almost certainly contains heteroskedasticity. In fact, if $\text{Var}(c_i|x_i) = \sigma_c^2$, $\text{Var}(d_i|x_i) = \sigma_d^2$, and $\text{Cov}(c_i,d_i|x_i) = 0$, then

$$\text{Var}(u_i|x_i) = \sigma_c^2 + \sigma_d^2 x_i^2, \tag{9.20}$$

and so there must be heteroskedasticity in u_i unless $\sigma_d^2 = 0$, which means $b_i = \beta$ for all i. We know how to account for heteroskedasticity of this kind. We can use OLS and compute heteroskedasticity-robust standard errors and test statistics, or we can estimate the variance function in (9.20) and apply weighted least squares. Of course, the latter strategy imposes homoskedasticity on the random intercept and slope, and so we would want to make a WLS analysis fully robust to violations of (9.20).

Because of equation (9.20), some authors like to view heteroskedasticity in regression models generally as arising from random slope coefficients. But we should remember that the form of (9.20) is special, and it does not allow for heteroskedasticity in a_i or b_i. We cannot convincingly distinguish between a random slope model, where the intercept and slope are independent of x_i, and a constant slope model with heteroskedasticity in a_i.

The treatment for multiple regression is similar. Generally, write

$$y_i = a_i + b_{i1}x_{i1} + b_{i2}x_{i2} + \cdots + b_{ik}x_{ik}. \tag{9.21}$$

Then, by writing $a_i = \alpha + c_i$ and $b_{ij} = \beta_j + d_{ij}$, we have

$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i, \tag{9.22}$$

where $u_i = c_i + d_{i1}x_{i1} + \cdots + d_{ik}x_{ik}$. If we maintain the mean independence assumptions $E(a_i|\mathbf{x}_i) = E(a_i)$ and $E(b_{ij}|\mathbf{x}_i) = E(b_{ij}), j = 1, \ldots, k$, then $E(y_i|\mathbf{x}_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, and so OLS using a random sample produces unbiased estimators of α and the β_j. As in the simple regression case, $\text{Var}(u_i|\mathbf{x}_i)$ is almost certainly heteroskedastic.

We can allow the b_{ij} to depend on observable explanatory variables as well as unobservables. For example, suppose with $k = 2$ the effect of x_{i2} depends on x_{i1}, and we write $b_{i2} = \beta_2 + \delta_1(x_{i1} - \mu_1) + d_{i2}$, where $\mu_1 = E(x_{i1})$. If we assume $E(d_{i2}|\mathbf{x}_i) = 0$ (and similarly for c_i and d_{i1}), then $E(y_i|x_{i1}, x_{i2}) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_1(x_{i1} - \mu_1)x_{i2}$, which means we have an interaction between x_{i1} and x_{i2}. Because we have subtracted the mean μ_1 from x_{i1}, β_2 is the APE of x_{i2}.

The bottom line of this section is that allowing for random slopes is fairly straightforward if the slopes are independent, or at least mean independent, of the explanatory variables. In addition, we can easily model the slopes as functions of the exogenous variables, which leads to models with squares and interactions. Of course, in Chapter 6 we discussed how such models can be useful without ever introducing the notion of a random slope. The random slopes specification provides a separate justification for such models. Estimation becomes considerably more difficult if the random intercept as well as some slopes are correlated with some of the regressors. We cover the problem of endogenous explanatory variables in Chapter 15.

9-4 Properties of OLS under Measurement Error

Sometimes, in economic applications, we cannot collect data on the variable that truly affects economic behavior. A good example is the marginal income tax rate facing a family that is trying to choose how much to contribute to charity in a given year. The marginal rate may be hard to obtain or summarize as a single number for all income levels. Instead, we might compute the average tax rate based on total income and tax payments.

When we use an imprecise measure of an economic variable in a regression model, then our model contains measurement error. In this section, we derive the consequences of measurement error for ordinary least squares estimation. OLS will be consistent under certain assumptions, but there are others under which it is inconsistent. In some of these cases, we can derive the size of the asymptotic bias.

As we will see, the measurement error problem has a similar statistical structure to the omitted variable–proxy variable problem discussed in the previous section, but they are conceptually different. In the proxy variable case, we are looking for a variable that is somehow associated with the unobserved variable. In the measurement error case, the variable that we do not observe has a well-defined, quantitative meaning (such as a marginal tax rate or annual income), but our recorded measures of it may contain error. For example, reported annual income is a measure of actual annual income, whereas IQ score is a proxy for ability.

Another important difference between the proxy variable and measurement error problems is that, in the latter case, often the mismeasured independent variable is the one of primary interest. In the proxy variable case, the partial effect of the omitted variable is rarely of central interest: we are usually concerned with the effects of the other independent variables.

Before we consider details, we should remember that measurement error is an issue only when the variables for which the econometrician can collect data differ from the variables that influence decisions by individuals, families, firms, and so on.

9-4a Measurement Error in the Dependent Variable

We begin with the case where only the dependent variable is measured with error. Let y^* denote the variable (in the population, as always) that we would like to explain. For example, y^* could be annual family savings. The regression model has the usual form

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \qquad [9.23]$$

and we assume it satisfies the Gauss-Markov assumptions. We let y represent the observable measure of y^*. In the savings case, y is reported annual savings. Unfortunately, families are not perfect in their reporting of annual family savings; it is easy to leave out categories or to overestimate the amount contributed to a fund. Generally, we can expect y and y^* to differ, at least for some subset of families in the population.

The **measurement error** (in the population) is defined as the difference between the observed value and the actual value:

$$e_0 = y - y^*. \qquad [9.24]$$

For a random draw i from the population, we can write $e_{i0} = y_i - y_i^*$, but the important thing is how the measurement error in the population is related to other factors. To obtain an estimable model, we write $y^* = y - e_0$, plug this into equation (9.23), and rearrange:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u + e_0. \qquad [9.25]$$

The error term in equation (9.25) is $u + e_0$. Because $y, x_1, x_2, \ldots, x_k$ are observed, we can estimate this model by OLS. In effect, we just ignore the fact that y is an imperfect measure of y^* and proceed as usual.

When does OLS with y in place of y^* produce consistent estimators of the β_j? Because the original model (9.23) satisfies the Gauss-Markov assumptions, u has zero mean and is uncorrelated with each x_j. It is only natural to assume that the measurement error has zero mean; if it does not, then we simply get a biased estimator of the intercept, β_0, which is rarely a cause for concern. Of much more importance is our assumption about the relationship between the measurement error, e_0, and the explanatory variables, x_j. The usual assumption is that the measurement error in y is statistically independent of each explanatory variable. If this is true, then the OLS estimators from (9.25) are unbiased and consistent. Further, the usual OLS inference procedures (t, F, and LM statistics) are valid.

If e_0 and u are uncorrelated, as is usually assumed, then $\text{Var}(u + e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2$. This means that measurement error in the dependent variable results in a larger error variance than when no error occurs; this, of course, results in larger variances of the OLS estimators. This is to be expected and there is nothing we can do about it (except collect better data). The bottom line is that, if the measurement error is uncorrelated with the independent variables, then OLS estimation has good properties.

EXAMPLE 9.5 Savings Function with Measurement Error

Consider a savings function

$$sav^* = \beta_0 + \beta_1 inc + \beta_2 size + \beta_3 educ + \beta_4 age + u,$$

but where actual savings (sav^*) may deviate from reported savings (sav). The question is whether the size of the measurement error in sav is systematically related to the other variables. It might be reasonable to assume that the measurement error is not correlated with inc, $size$, $educ$, and age. On the other hand, we might think that families with higher incomes, or more education, report their savings more accurately. We can never know whether the measurement error is correlated with inc or $educ$, unless we can collect data on sav^*; then, the measurement error can be computed for each observation as $e_{i0} = sav_i - sav_i^*$.

When the dependent variable is in logarithmic form, so that $\log(y^*)$ is the dependent variable, it is natural for the measurement error equation to be of the form

$$\log(y) = \log(y^*) + e_0. \qquad [9.26]$$

This follows from a **multiplicative measurement error** for y: $y = y^* a_0$, where $a_0 > 0$ and $e_0 = \log(a_0)$.

EXAMPLE 9.6 Measurement Error in Scrap Rates

In Section 7-6, we discussed an example in which we wanted to determine whether job training grants reduce the scrap rate in manufacturing firms. We certainly might think the scrap rate reported by firms is measured with error. (In fact, most firms in the sample do not even report a scrap rate.) In a simple regression framework, this is captured by

$$\log(scrap^*) = \beta_0 + \beta_1 grant + u,$$

where $scrap^*$ is the true scrap rate and $grant$ is the dummy variable indicating whether a firm received a grant. The measurement error equation is

$$\log(scrap) = \log(scrap^*) + e_0.$$

Is the measurement error, e_0, independent of whether the firm receives a grant? A cynical person might think that a firm receiving a grant is more likely to underreport its scrap rate in order to make the grant look effective. If this happens, then, in the estimable equation,

$$\log(scrap) = \beta_0 + \beta_1 grant + u + e_0,$$

the error $u + e_0$ is negatively correlated with *grant*. This would produce a downward bias in β_1, which would tend to make the training program look more effective than it actually was. (Remember, a more negative β_1 means the program was more effective, because increased worker productivity is associated with a lower scrap rate.)

The bottom line of this subsection is that measurement error in the dependent variable *can* cause biases in OLS if it is systematically related to one or more of the explanatory variables. If the measurement error is just a random reporting error that is independent of the explanatory variables, as is often assumed, then OLS is perfectly appropriate.

9-4b Measurement Error in an Explanatory Variable

Traditionally, measurement error in an explanatory variable has been considered a much more important problem than measurement error in the dependent variable. In this subsection, we will see why this is the case.

We begin with the simple regression model

$$y = \beta_0 + \beta_1 x_1^* + u, \qquad [9.27]$$

and we assume that this satisfies at least the first four Gauss-Markov assumptions. This means that estimation of (9.27) by OLS would produce unbiased and consistent estimators of β_0 and β_1. The problem is that x_1^* is not observed. Instead, we have a measure of x_1^*; call it x_1. For example, x_1^* could be actual income and x_1 could be reported income.

The measurement error in the population is simply

$$e_1 = x_1 - x_1^*, \qquad [9.28]$$

and this can be positive, negative, or zero. We assume that the *average* measurement error in the population is zero: $\mathrm{E}(e_1) = 0$. This is natural, and, in any case, it does not affect the important conclusions that follow. A maintained assumption in what follows is that u is uncorrelated with x_1^* and x_1. In conditional expectation terms, we can write this as $\mathrm{E}(y|x_1^*, x_1) = \mathrm{E}(y|x_1^*)$, which just says that x_1 does not affect y after x_1^* has been controlled for. We used the same assumption in the proxy variable case, and it is not controversial; it holds almost by definition.

We want to know the properties of OLS if we simply replace x_1^* with x_1 and run the regression of y on x_1. They depend crucially on the assumptions we make about the measurement error. Two assumptions have been the focus in econometrics literature, and they both represent polar extremes. The first assumption is that e_1 is uncorrelated with the *observed* measure, x_1:

$$\mathrm{Cov}(x_1, e_1) = 0. \qquad [9.29]$$

From the relationship in (9.28), if assumption (9.29) is true, then e_1 must be correlated with the unobserved variable x_1^*. To determine the properties of OLS in this case, we write $x_1^* = x_1 - e_1$ and plug this into equation (9.27):

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1). \qquad [9.30]$$

Because we have assumed that u and e_1 both have zero mean and are uncorrelated with x_1, $u - \beta_1 e_1$ has zero mean and is uncorrelated with x_1. It follows that OLS estimation with x_1 in place of x_1^*

produces a consistent estimator of β_1 (and also β_0). Because u is uncorrelated with e_1, the variance of the error in (9.30) is $\text{Var}(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2\sigma_{e_1}^2$. Thus, except when $\beta_1 = 0$, measurement error increases the error variance. But this does not affect any of the OLS properties (except that the variances of the $\hat{\beta}_j$ will be larger than if we observe x_1^* directly).

The assumption that e_1 is uncorrelated with x_1 is analogous to the proxy variable assumption we made in Section 9-2. Because this assumption implies that OLS has all of its nice properties, this is not usually what econometricians have in mind when they refer to measurement error in an explanatory variable. The **classical errors-in-variables (CEV)** assumption is that the measurement error is uncorrelated with the *unobserved* explanatory variable:

$$\text{Cov}(x_1^*, e_1) = 0. \qquad [9.31]$$

This assumption comes from writing the observed measure as the sum of the true explanatory variable and the measurement error,

$$x_1 = x_1^* + e_1,$$

and then assuming the two components of x_1 are uncorrelated. (This has nothing to do with assumptions about u; we always maintain that u is uncorrelated with x_1^* and x_1, and therefore with e_1.)

If assumption (9.31) holds, then x_1 and e_1 *must* be correlated:

$$\text{Cov}(x_1, e_1) = \text{E}(x_1 e_1) = \text{E}(x_1^* e_1) + \text{E}(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2. \qquad [9.32]$$

Thus, the covariance between x_1 and e_1 is equal to the variance of the measurement error under the CEV assumption.

Referring to equation (9.30), we can see that correlation between x_1 and e_1 is going to cause problems. Because u and x_1 are uncorrelated, the covariance between x_1 and the composite error $u - \beta_1 e_1$ is

$$\text{Cov}(x_1, u - \beta_1 e_1) = -\beta_1 \text{Cov}(x_1, e_1) = -\beta_1 \sigma_{e_1}^2.$$

Thus, in the CEV case, the OLS regression of y on x_1 gives a biased and inconsistent estimator.

Using the asymptotic results in Chapter 5, we can determine the amount of inconsistency in OLS. The probability limit of $\hat{\beta}_1$ is β_1 plus the ratio of the covariance between x_1 and $u - \beta_1 e_1$ and the variance of x_1:

$$\begin{aligned} \text{plim}(\hat{\beta}_1) &= \beta_1 + \frac{\text{Cov}(x_1, u - \beta_1 e_1)}{\text{Var}(x_1)} \\ &= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1}^{2*} + \sigma_{e_1}^2} \\ &= \beta_1 \left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1}^{2*} + \sigma_{e_1}^2}\right) \\ &= \beta_1 \left(\frac{\sigma_{x_1}^{2*}}{\sigma_{x_1}^{2*} + \sigma_{e_1}^2}\right), \end{aligned} \qquad [9.33]$$

where we have used the fact that $\text{Var}(x_1) = \text{Var}(x_1^*) + \text{Var}(e_1)$.

Equation (9.33) is very interesting. The term multiplying β_1, which is the ratio $\text{Var}(x_1^*)/\text{Var}(x_1)$, is always less than one [an implication of the CEV assumption (9.31)]. Thus, $\text{plim}(\hat{\beta}_1)$ is always closer to zero than is β_1. This is called the **attenuation bias** in OLS due to CEV: on average (or in large samples), the estimated OLS effect will be *attenuated*. In particular, if β_1 is positive, $\hat{\beta}_1$ will tend to underestimate β_1. This is an important conclusion, but it relies on the CEV setup.

If the variance of x_1^* is large relative to the variance in the measurement error, then the inconsistency in OLS will be small. This is because $\text{Var}(x_1^*)/\text{Var}(x_1)$ will be close to unity when $\sigma_{x_1^*}^2/\sigma_{e_1}^2$ is

large. Therefore, depending on how much variation there is in x_1^* relative to e_1, measurement error need not cause large biases.

Things are more complicated when we add more explanatory variables. For illustration, consider the model

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + \beta_2 x_3 + u, \qquad [9.34]$$

where the first of the three explanatory variables is measured with error. We make the natural assumption that u is uncorrelated with x_1^*, x_2, x_3, and x_1. Again, the crucial assumption concerns the measurement error e_1. In almost all cases, e_1 is assumed to be uncorrelated with x_2 and x_3—the explanatory variables not measured with error. The key issue is whether e_1 is uncorrelated with x_1. If it is, then the OLS regression of y on x_1, x_2, and x_3 produces consistent estimators. This is easily seen by writing

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + u - \beta_1 e_1, \qquad [9.35]$$

where u and e_1 are both uncorrelated with all the explanatory variables.

Under the CEV assumption in (9.31), OLS will be biased and inconsistent, because e_1 is correlated with x_1 in equation (9.35). Remember, this means that, in general, *all* OLS estimators will be biased, not just $\hat{\beta}_1$. What about the attenuation bias derived in equation (9.33)? It turns out that there is still an attenuation bias for estimating β_1: it can be shown that

$$\text{plim}(\hat{\beta}_1) = \beta_1\left(\frac{\sigma_{r_1}^{2*}}{\sigma_{r_1}^{2*} + \sigma_{e_1}^2}\right), \qquad [9.36]$$

where r_1^* is the population error in the equation $x_1^* = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + r_1^*$. Formula (9.36) also works in the general k variable case when x_1 is the only mismeasured variable.

Things are less clear-cut for estimating the β_j on the variables not measured with error. In the special case that x_1^* is uncorrelated with x_2 and x_3, $\hat{\beta}_2$ and $\hat{\beta}_3$ are consistent. But this is rare in practice. Generally, measurement error in a single variable causes inconsistency in all estimators. Unfortunately, the sizes, and even the directions of the biases, are not easily derived.

EXAMPLE 9.7 GPA Equation with Measurement Error

Consider the problem of estimating the effect of family income on college grade point average, after controlling for *hsGPA* (high school grade point average) and *SAT* (scholastic aptitude test). It could be that, though family income is important for performance before college, it has no direct effect on college performance. To test this, we might postulate the model

$$colGPA = \beta_0 + \beta_1 faminc^* + \beta_2 hsGPA + \beta_3 SAT + u,$$

where $faminc^*$ is actual annual family income. (This might appear in logarithmic form, but for the sake of illustration we leave it in level form.) Precise data on *colGPA*, *hsGPA*, and *SAT* are relatively easy to obtain. But family income, especially as reported by students, could be easily mismeasured. If $faminc = faminc^* + e_1$ and the CEV assumptions hold, then using reported family income in place of actual family income will bias the OLS estimator of β_1 toward zero. One consequence of the downward bias is that a test of H_0: $\beta_1 = 0$ will have less chance of detecting $\beta_1 > 0$.

Of course, measurement error can be present in more than one explanatory variable, or in some explanatory variables and the dependent variable. As we discussed earlier, any measurement error in the dependent variable is usually assumed to be uncorrelated with all the explanatory variables, whether it is observed or not. Deriving the bias in the OLS estimators under extensions of the CEV assumptions is complicated and does not lead to clear results.

In some cases, it is clear that the CEV assumption in (9.31) cannot be true. Consider a variant on Example 9.7:

$$colGPA = \beta_0 + \beta_1 smoked^* + \beta_2 hsGPA + \beta_3 SAT + u,$$

where $smoked^*$ is the actual number of times a student smoked marijuana in the last 30 days. The variable *smoked* is the answer to this question: On how many separate occasions did you smoke marijuana in the last 30 days? Suppose we postulate the standard measurement error model

$$smoked = smoked^* + e_1.$$

Even if we assume that students try to report the truth, the CEV assumption is unlikely to hold. People who do not smoke marijuana at all—so that $smoked^* = 0$—are likely to report $smoked = 0$, so the measurement error is probably zero for students who never smoke marijuana. When $smoked^* > 0$, it is much more likely that the student miscounts how many times he or she smoked marijuana in the last 30 days. This means that the measurement error e_1 and the *actual* number of times smoked, $smoked^*$, are correlated, which violates the CEV assumption in (9.31). Unfortunately, deriving the implications of measurement error that do not satisfy (9.29) or (9.31) is difficult and beyond the scope of this text.

GOING FURTHER 9.3

Let $educ^*$ be actual amount of schooling, measured in years (which can be a noninteger), and let *educ* be reported highest grade completed. Do you think *educ* and $educ^*$ are related by the CEV model?

Before leaving this section, we emphasize that the CEV assumption (9.31), while more believable than assumption (9.29), is still a strong assumption. The truth is probably somewhere in between, and if e_1 is correlated with both x_1^* and x_1, OLS is inconsistent. This raises an important question: must we live with inconsistent estimators under CEV, or other kinds of measurement error that are correlated with x_1? Fortunately, the answer is no. Chapter 15 shows how, under certain assumptions, the parameters can be consistently estimated in the presence of general measurement error. We postpone this discussion until later because it requires us to leave the realm of OLS estimation. (See Problem 7 for how multiple measures can be used to reduce the attenuation bias.)

9-5 Missing Data, Nonrandom Samples, and Outlying Observations

The measurement error problem discussed in the previous section can be viewed as a data problem: we cannot obtain data on the variables of interest. Further, under the CEV model, the composite error term is correlated with the mismeasured independent variable, violating the Gauss-Markov assumptions.

Another data problem we discussed frequently in earlier chapters is multicollinearity among the explanatory variables. Remember that correlation among the explanatory variables does not violate any assumptions. When two independent variables are highly correlated, it can be difficult to estimate the partial effect of each. But this is properly reflected in the usual OLS statistics.

In this section, we provide an introduction to data problems that can violate the random sampling assumption, MLR.2. We can isolate cases in which nonrandom sampling has no practical effect on OLS. In other cases, nonrandom sampling causes the OLS estimators to be biased and inconsistent. A more complete treatment that establishes several of the claims made here is given in Chapter 17.

9-5a Missing Data

The **missing data** problem can arise in a variety of forms. Often, we collect a random sample of people, schools, cities, and so on, and then discover later that information is missing on some key variables for several units in the sample. For example, in the data set BWGHT, 196 of the 1,388

observations have no information on father's education. In the data set on median starting law school salaries, LAWSCH85, 6 of the 156 schools have no reported information on median LSAT scores for the entering class; other variables are also missing for some of the law schools.

If data are missing for an observation on either the dependent variable or one of the independent variables, then the observation cannot be used in a standard multiple regression analysis. In fact, provided missing data have been properly indicated, all modern regression packages keep track of missing data and simply ignore observations when computing a regression. We saw this explicitly in the birth weight scenario in Example 4.9, when 197 observations were dropped due to missing information on parents' education.

In the literature on missing data, an estimator that uses only observations with a complete set of data on y and $x_1, \ldots, x_k$ is called a **complete cases estimator**; as mentioned earlier, this estimator is computed as the default for OLS (and all estimators covered later in the text). Other than reducing the sample size, are there any *statistical* consequences of using the OLS estimator and ignoring the missing data? If, in the language of the missing data literature [see, for example, Little and Rubin (2002, Chapter 1)] the data are **missing completely at random** (sometimes called **MCAR**), then missing data cause no statistical problems. The MCAR assumption implies that the reason the data are missing is independent, in a statistical sense, of both the observed and unobserved factors affecting y. In effect, we can still assume that the data have been obtained by random sampling from the population, so that Assumption MLR.2 continues to hold.

When MCAR holds, there are ways to use partial information obtained from units that are dropped from the complete case estimation. Unfortunately, some simple strategies produce consistent estimators only under strong assumptions—in addition to MCAR. As illustration, suppose that for a multiple regression model data are always available for y and $x_1, x_2, \ldots, x_{k-1}$ but are sometimes missing for the explanatory variable x_k. A common "solution" is to create two new variables. For a unit i, the first variable, say Z_{ik}, is defined to be x_{ik} when x_{ik} is observed, and zero otherwise. The second variable is a "missing data indicator," say m_{ik}, which equals one when x_{ik} is missing and equals zero when x_{ik} is observed. Having defined these two variables, all of the units are used in the regression

$$y_i \text{ on } x_{i1}, x_{i2}, \ldots, x_{i,k-1}, Z_{ik}, m_{ik}, \; i = 1, \ldots, n.$$

It is easy to understand the appeal of this procedure, which we call the **missing indicator method (MIM)**. Suppose that the original sample size is $n = 1{,}000$ but x_{ik} is missing for 30% of the cases. The complete cases estimator would use only 700 observations while the MIM regression would be based on all 1,000 cases. Unfortunately, the gain in observations is largely illusory, as the MIM estimator only has good statistical properties under strong assumptions. In particular, in addition to MCAR, consistency essentially requires that x_k is uncorrelated with the other explanatory variables, $x_1, x_2, \ldots, x_{k-1}$, as discussed in Jones (1996) and expanded on in Abrevaya and Donald (2018). Of course, it is difficult to know if the bias and inconsistency in MIM is practically important, but we have no way of generally knowing. One thing we can be sure of is that it is a very poor idea to omit m_{ik} from the regression, as that is the same as setting x_{ik} equal to zero whenever it is missing. Problem 9.10 works through how MCAR is sufficient for consistency in the simple regression model. The reader is referred to Abrevaya and Donald (2018) to see why MCAR is not sufficient when other regressors are included. In addition, Abrevaya and Donald (2018) discuss more robust ways to include information when some variables have missing data. The methods are too advanced for the scope of this text.

An important consequence of the previous discussion is that MIM is substantially less robust than the complete case estimator in the sense that the MIM approach requires much stronger assumptions for consistency. As we will see in the next subsection, the complete cases estimator turns out to be consistent even when the reason the data are missing is a function of $(x_1, x_2, \ldots, x_k)$, something explicitly ruled out by MCAR. Plus, the complete cases estimator puts no restrictions on the correlations among $(x_1, x_2, \ldots, x_k)$.

There are more complicated schemes for using partial information that are based on filling in the missing data, but these are beyond the scope of this text. The reader is referred to Little and Rubin (2002) and Abrevaya and Donald (2018).

9-5b Nonrandom Samples

The MCAR assumption ensures that units for which we observe a full set of data are not systematically different from units for which some variables are missing. Unfortunately, MCAR is often unrealistic. An example of a missing data mechanism that does not satisfy MCAR can be gotten by looking at the data set CARD, where the measure of IQ is missing for 949 men. If the probability that the IQ score is missing is, say, higher for men with lower IQ scores, the mechanism violates MCAR. For example, in the birth weight data set, what if the probability that education is missing is higher for those people with lower than average levels of education? Or, in Section 9-2, we used a wage data set that included IQ scores. This data set was constructed by omitting several people from the sample for whom IQ scores were not available. If obtaining an IQ score is easier for those with higher IQs, the sample is not representative of the population. The random sampling assumption MLR.2 is violated, and we must worry about these consequences for OLS estimation.

Fortunately, certain types of nonrandom sampling do *not* cause bias or inconsistency in OLS. Under the Gauss-Markov assumptions (but without MLR.2), it turns out that the sample can be chosen on the basis of the *independent* variables without causing any statistical problems. This is called *sample selection based on the independent variables*, and it is an example of **exogenous sample selection**.

In the statistics literature, exogenous sample selection due to missing data is often called **missing at random (MAR)**, which is not a particularly good label because the probability of missing data is allowed to depend on the explanatory variables. The word "random" would seem to connote that missingness cannot depend systematically on anything, but that is actually the intention of the phrase "completely at random." In other words, MAR requires that missingness is unrelated to u but allows it to depend on $(x_1, x_2, \ldots, x_k)$, whereas MCAR means the missingness is unrelated to $(x_1, x_2, \ldots, x_k)$ and u. See Little and Rubin (2002, Chapter 1) for further discussion.

To illustrate exogenously missing data, suppose that we are estimating a saving function, where annual saving depends on income, age, family size, and some unobserved factors, u. A simple model is

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + u. \qquad [9.37]$$

Suppose that our data set was based on a survey of people over 35 years of age, thereby leaving us with a nonrandom sample of all adults. While this is not ideal, we can still get unbiased and consistent estimators of the parameters in the population model (9.37), using the nonrandom sample. We will not show this formally here, but the reason OLS on the nonrandom sample is unbiased is that the regression function $\mathrm{E}(saving|income,age,size)$ is the same for any subset of the population described by *income*, *age*, or *size*. Provided there is enough variation in the independent variables in the subpopulation, selection on the basis of the independent variables is not a serious problem, other than that it results in smaller sample sizes.

In the IQ example just mentioned, things are not so clear-cut, because no fixed rule based on IQ is used to include someone in the sample. Rather, the *probability* of being in the sample increases with IQ. If the other factors determining selection into the sample are independent of the error term in the wage equation, then we have another case of exogenous sample selection, and OLS using the selected sample will have all of its desirable properties under the other Gauss-Markov assumptions.

The situation is much different when selection is based on the dependent variable, y, which is called *sample selection based on the dependent variable* and is an example of **endogenous sample selection**. If the sample is based on whether the dependent variable is above or below a given value, bias always occurs in OLS in estimating the population model. For example, suppose we wish to

estimate the relationship between individual wealth and several other factors in the population of all adults:

$$wealth = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + u. \quad [9.38]$$

Suppose that only people with wealth below $250,000 are included in the sample. This is a nonrandom sample from the population of interest, and it is based on the value of the dependent variable. Using a sample of people with wealth below $250,000 will result in biased and inconsistent estimators of the parameters in (9.32). Briefly, this occurs because the population regression $E(wealth|educ,exper,age)$ is not the same as the expected value conditional on *wealth* being less than $250,000.

Other sampling schemes lead to **nonrandom samples** from the population, usually intentionally. A common method of data collection is **stratified sampling**, in which the population is divided into nonoverlapping, exhaustive groups, or strata. Then, some groups are sampled more frequently than is dictated by their population representation, and some groups are sampled less frequently. For example, some surveys purposely oversample minority groups or low-income groups. Whether special methods are needed again hinges on whether the stratification is exogenous (based on exogenous explanatory variables) or endogenous (based on the dependent variable). Suppose that a survey of military personnel oversampled women because the initial interest was in studying the factors that determine pay for women in the military. (Oversampling a group that is relatively small in the population is common in collecting stratified samples.) Provided men were sampled as well, we can use OLS on the stratified sample to estimate any gender differential, along with the returns to education and experience for all military personnel. (We might be willing to assume that the returns to education and experience are not gender specific.) OLS is unbiased and consistent because the stratification is with respect to an explanatory variable, namely, gender.

If, instead, the survey oversampled lower-paid military personnel, then OLS using the stratified sample does not consistently estimate the parameters of the military wage equation because the stratification is endogenous. In such cases, special econometric methods are needed [see Wooldridge (2010, Chapter 19)].

Stratified sampling is a fairly obvious form of nonrandom sampling. Other sample selection issues are more subtle. For instance, in several previous examples, we have estimated the effects of various variables, particularly education and experience, on hourly wage. The data set WAGE1 that we have used throughout is essentially a random sample of *working* individuals. Labor economists are often interested in estimating the effect of, say, education on the wage *offer*. The idea is this: every person of working age faces an hourly wage offer, and he or she can either work at that wage or not work. For someone who does work, the wage offer is just the wage earned. For people who do not work, we usually cannot observe the wage offer. Now, because the wage offer equation

$$\log(wage^o) = \beta_0 + \beta_1 educ + \beta_2 exper + u \quad [9.39]$$

represents the population of all working-age people, we cannot estimate it using a random sample from this population; instead, we have data on the wage offer only for working people (although we can get data on *educ* and *exper* for nonworking people). If we use a random sample of working people to estimate (9.39), will we get unbiased estimators? This case is not clear-cut. Because the sample is selected based on someone's decision to work (as opposed to the size of the wage offer), this is not like the previous case. However, because the decision to work might be related to unobserved factors that affect the wage offer, selection might be endogenous, and this can result in a sample selection bias in the OLS estimators. We will cover methods that can be used to test and correct for sample selection bias in Chapter 17.

GOING FURTHER 9.4

Suppose we are interested in the effects of campaign expenditures by incumbents on voter support. Some incumbents choose not to run for reelection. If we can only collect voting and spending outcomes on incumbents that actually do run, is there likely to be endogenous sample selection?

9-5c Outliers and Influential Observations

In some applications, especially, but not only, with small data sets, the OLS estimates are sensitive to the inclusion of one or several observations. A complete treatment of **outliers** and **influential observations** is beyond the scope of this book, because a formal development requires matrix algebra. Loosely speaking, an observation is an influential observation if dropping it from the analysis changes the key OLS estimates by a practically "large" amount. The notion of an outlier is also a bit vague, because it requires comparing values of the variables for one observation with those for the remaining sample. Nevertheless, one wants to be on the lookout for "unusual" observations because they can greatly affect the OLS estimates.

OLS is susceptible to outlying observations because it minimizes the sum of squared residuals: large residuals (positive or negative) receive a lot of weight in the least squares minimization problem. If the estimates change by a practically large amount when we slightly modify our sample, we should be concerned.

When statisticians and econometricians study the problem of outliers theoretically, sometimes the data are viewed as being from a random sample from a given population—albeit with an unusual distribution that can result in extreme values—and sometimes the outliers are assumed to come from a different population. From a practical perspective, outlying observations can occur for two reasons. The easiest case to deal with is when a mistake has been made in entering the data. Adding extra zeros to a number or misplacing a decimal point can throw off the OLS estimates, especially in small sample sizes. It is always a good idea to compute summary statistics, especially minimums and maximums, in order to catch mistakes in data entry. Unfortunately, incorrect entries are not always obvious.

Outliers can also arise when sampling from a small population if one or several members of the population are very different in some relevant aspect from the rest of the population. The decision to keep or drop such observations in a regression analysis can be a difficult one, and the statistical properties of the resulting estimators are complicated. Outlying observations can provide important information by increasing the variation in the explanatory variables (which reduces standard errors). But OLS results should probably be reported with and without outlying observations in cases where one or several data points substantially change the results.

EXAMPLE 9.8 R&D Intensity and Firm Size

Suppose that R&D expenditures as a percentage of sales (*rdintens*) are related to *sales* (in millions) and profits as a percentage of sales (*profmarg*):

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 profmarg + u. \qquad [9.40]$$

The OLS equation using data on 32 chemical companies in RDCHEM is

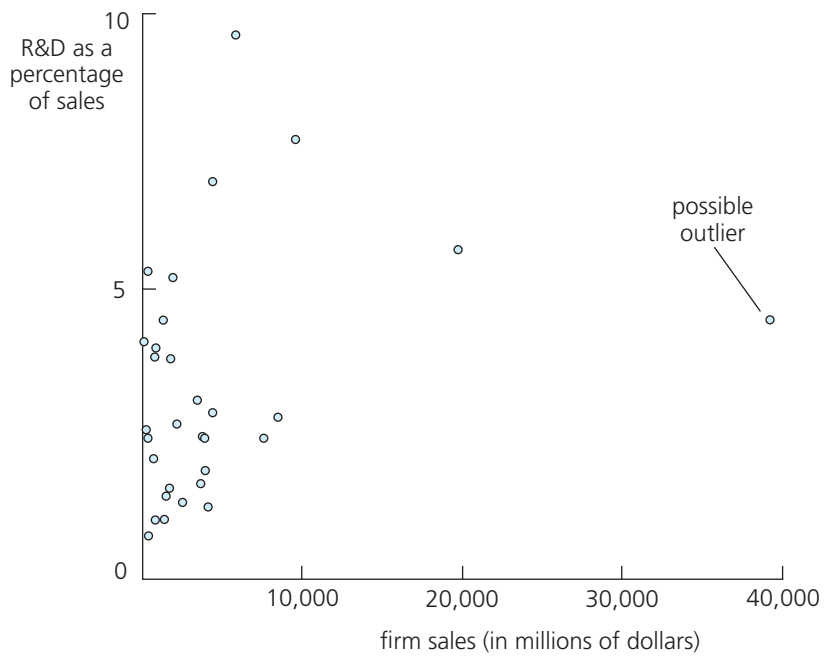
$$\widehat{rdintens} = 2.625 + .000053\, sales + .0446\, profmarg$$
$$(0.586)\ \ (.000044)\qquad\ (.0462)$$
$$n = 32, R^2 = .0761, \overline{R}^2 = .0124.$$

Neither *sales* nor *profmarg* is statistically significant at even the 10% level in this regression.

Of the 32 firms, 31 have annual sales less than $20 billion. One firm has annual sales of almost $40 billion. Figure 9.1 shows how far this firm is from the rest of the sample. In terms of sales, this firm is over twice as large as every other firm, so it might be a good idea to estimate the model without it. When we do this, we obtain

$$\widehat{rdintens} = 2.297 + .000186\, sales + .0478\, profmarg$$
$$(0.592)\ \ (.000084)\qquad\ (.0445)$$
$$n = 31, R^2 = .1728, \overline{R}^2 = .1137.$$

FIGURE 9.1 Scatterplot of R&D intensity against firm sales.



When the largest firm is dropped from the regression, the coefficient on *sales* more than triples, and it now has a t statistic over two. Using the sample of smaller firms, we would conclude that there is a statistically significant positive effect between R&D intensity and firm size. The profit margin is still not significant, and its coefficient has not changed by much.

Sometimes, outliers are defined by the size of the residual in an OLS regression, where all of the observations are used. Generally, this is *not* a good idea because the OLS estimates adjust to make the sum of squared residuals as small as possible. In the previous example, including the largest firm flattened the OLS regression line considerably, which made the residual for that estimation not especially large. In fact, the residual for the largest firm is −1.62 when all 32 observations are used. This value of the residual is not even one estimated standard deviation, $\hat{\sigma} = 1.82$, from the mean of the residuals, which is zero by construction.

Studentized residuals are obtained from the original OLS residuals by dividing them by an estimate of their standard deviation (conditional on the explanatory variables in the sample). The formula for the studentized residuals relies on matrix algebra, but it turns out there is a simple trick to compute a studentized residual for any observation. Namely, define a dummy variable equal to one for that observation—say, observation h—and then include the dummy variable in the regression (using all observations) along with the other explanatory variables. The coefficient on the dummy variable has a useful interpretation: it is the residual for observation h computed from the regression line using only the *other* observations. Therefore, the dummy's coefficient can be used to see how far off the observation is from the regression line obtained without using that observation. Even better, the t statistic on the dummy variable is equal to the studentized residual for observation h. Under the classical linear model assumptions, this t statistic has a t_{n-k-2} distribution. Therefore, a large value of the t statistic (in absolute value) implies a large residual relative to its estimated standard deviation.

For Example 9.8, if we define a dummy variable for the largest firm (observation 10 in the data file), and include it as an additional regressor, its coefficient is −6.57, verifying that the observation for the largest firm is very far from the regression line obtained using the other observations. However, when studentized, the residual is only −1.82. While this is a marginally significant t statistic (two-sided p-value = .08), it is not close to being the largest studentized residual in the sample. If we use the same method for the observation with the highest value of *rdintens*—the first observation, with *rdintens* ≈ 9.42—the coefficient on the dummy variable is 6.72 with a t statistic of 4.56. Therefore, by this measure, the first observation is more of an outlier than the tenth. Yet dropping the first observation changes the coefficient on *sales* by only a small amount (to about .000051 from .000053), although the coefficient on *profmarg* becomes larger and statistically significant. So, is the first observation an "outlier" too? These calculations show the conundrum one can enter when trying to determine observations that should be excluded from a regression analysis, even when the data set is small. Unfortunately, the size of the studentized residual need not correspond to how influential an observation is for the OLS slope estimates, and certainly not for all of them at once.

A general problem with using studentized residuals is that, in effect, all other observations are used to estimate the regression line to compute the residual for a particular observation. In other words, when the studentized residual is obtained for the first observation, the tenth observation has been used in estimating the intercept and slope. Given how flat the regression line is with the largest firm (tenth observation) included, it is not too surprising that the first observation, with its high value of *rdintens*, is far off the regression line.

Of course, we can add two dummy variables at the same time—one for the first observation and one for the tenth—which has the effect of using only the remaining 30 observations to estimate the regression line. If we estimate the equation without the first and tenth observations, the results are

$$\widehat{rdintens} = 1.939 + .000160\, sales + .0701\, profmarg$$
$$(0.459)\ \ (.00065)\qquad\quad (.0343)$$
$$n = 30, R^2 = .2711, \overline{R}^2 = .2171.$$

The coefficient on the dummy for the first observation is 6.47 $(t = 4.58)$, and for the tenth observation it is −5.41 $(t = -1.95)$. Notice that the coefficients on the *sales* and *profmarg* are both statistically significant, the latter at just about the 5% level against a two-sided alternative (p-value = .051). Even in this regression there are still two observations with studentized residuals greater than two (corresponding to the two remaining observations with R&D intensities above six).

Certain functional forms are less sensitive to outlying observations. In Section 6-2 we mentioned that, for most economic variables, the logarithmic transformation significantly narrows the range of the data and also yields functional forms—such as constant elasticity models—that can explain a broader range of data.

EXAMPLE 9.9 R&D Intensity

We can test whether R&D intensity increases with firm size by starting with the model

$$rd = sales^{\beta_1}\exp(\beta_0 + \beta_2 profmarg + u). \qquad [9.41]$$

Then, holding other factors fixed, R&D intensity increases with *sales* if and only if $\beta_1 > 1$. Taking the log of (9.41) gives

$$\log(rd) = \beta_0 + \beta_1 \log(sales) + \beta_2 profmarg + u. \qquad [9.42]$$

When we use all 32 firms, the regression equation is

$$\widehat{\log(rd)} = -4.378 + 1.084 \log(sales) + .0217\, profmarg,$$
$$\quad\quad (.468) \quad (.060) \quad\quad\quad (.0128)$$
$$n = 32, R^2 = .9180, \bar{R}^2 = .9123,$$

while dropping the largest firm gives

$$\widehat{\log(rd)} = -4.404 + 1.088 \log(sales) + .0218\, profmarg,$$
$$\quad\quad (.511) \quad (.067) \quad\quad\quad (.0130)$$
$$n = 31, R^2 = .9037, \bar{R}^2 = .8968.$$

Practically, these results are the same. In neither case do we reject the null H_0: $\beta_1 = 1$ against H_1: $\beta_1 > 1$. (Why?)

In some cases, certain observations are suspected at the outset of being fundamentally different from the rest of the sample. This often happens when we use data at very aggregated levels, such as the city, county, or state level. The following is an example.

EXAMPLE 9.10 State Infant Mortality Rates

Data on infant mortality, per capita income, and measures of health care can be obtained at the state level from the *Statistical Abstract of the United States*. We will provide a fairly simple analysis here just to illustrate the effect of outliers. The data are for the year 1990, and we have all 50 states in the United States, plus the District of Columbia (D.C.). The variable *infmort* is number of deaths within the first year per 1,000 live births, *pcinc* is per capita income, *physic* is physicians per 100,000 members of the civilian population, and *popul* is the population (in thousands). The data are contained in INFMRT. We include all independent variables in logarithmic form:

$$\widehat{infmort} = 33.86 - 4.68 \log(pcinc) + 4.15 \log(physic)$$
$$\quad\quad (20.43) \quad (2.60) \quad\quad\quad (1.51)$$
$$- .088 \log(popul) \quad\quad [9.43]$$
$$\quad\quad (.287)$$
$$n = 51, R^2 = .139, \bar{R}^2 = .084.$$

Higher per capita income is estimated to lower infant mortality, an expected result. But more physicians per capita is associated with *higher* infant mortality rates, something that is counterintuitive. Infant mortality rates do not appear to be related to population size.

The District of Columbia is unusual in that it has pockets of extreme poverty and great wealth in a small area. In fact, the infant mortality rate for D.C. in 1990 was 20.7, compared with 12.4 for the highest state. It also has 615 physicians per 100,000 of the civilian population, compared with 337 for the highest state. The high number of physicians coupled with the high infant mortality rate in D.C. could certainly influence the results. If we drop D.C. from the regression, we obtain

$$\widehat{infmort} = 23.95 - .57 \log(pcinc) - 2.74 \log(physic)$$
$$\quad\quad (12.42) (1.64) \quad\quad\quad (1.19)$$
$$+ .629 \log(popul) \quad\quad [9.44]$$
$$\quad\quad (.191)$$
$$n = 50, R^2 = .273, \bar{R}^2 = .226.$$

We now find that more physicians per capita lowers infant mortality, and the estimate is statistically different from zero at the 5% level. The effect of per capita income has fallen sharply and is no longer statistically significant. In equation (9.44), infant mortality rates are higher in more populous states, and the relationship is very statistically significant. Also, much more variation in *infmort* is explained when D.C. is dropped from the regression. Clearly, D.C. had substantial influence on the initial estimates, and we would probably leave it out of any further analysis.

As Example 9.8 demonstrates, inspecting observations in trying to determine which are outliers, and even which ones have substantial influence on the OLS estimates, is a difficult endeavor. More advanced treatments allow more formal approaches to determine which observations are likely to be influential observations. Using matrix algebra, Belsley, Kuh, and Welsh (1980) define the *leverage* of an observation, which formalizes the notion that an observation has a large or small influence on the OLS estimates. These authors also provide a more in-depth discussion of standardized and studentized residuals.

9-6 Least Absolute Deviations Estimation

Rather than trying to determine which observations, if any, have undue influence on the OLS estimates, a different approach to guarding against outliers is to use an estimation method that is less sensitive to outliers than OLS. One such method, which has become popular among applied econometricians, is called **least absolute deviations (LAD)**. The LAD estimators of the β_j in a linear model minimize the sum of the absolute values of the residuals,

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} |y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik}|. \qquad [9.45]$$

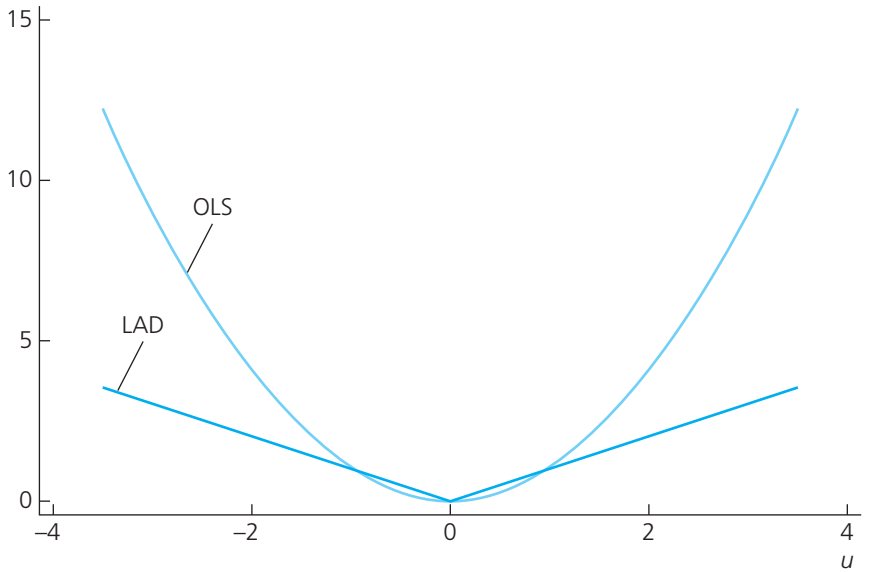
Unlike OLS, which minimizes the sum of squared residuals, the LAD estimates are not available in closed form—that is, we cannot write down formulas for them. In fact, historically, solving the problem in equation (9.45) was computationally difficult, especially with large sample sizes and many explanatory variables. But with the vast improvements in computational speed over the past two decades, LAD estimates are fairly easy to obtain even for large data sets.

Figure 9.2 shows the OLS and LAD objective functions. The LAD objective function is linear on either side of zero, so that if, say, a positive residual increases by one unit, the LAD objective function increases by one unit. By contrast, the OLS objective function gives increasing importance to large residuals, and this makes OLS more sensitive to outlying observations.

Because LAD does not give increasing weight to larger residuals, it is much less sensitive to changes in the extreme values of the data than OLS. In fact, it is known that LAD is designed to estimate the parameters of the **conditional median** of y given $x_1, x_2, \ldots, x_k$ rather than the conditional mean. Because the median is not affected by large changes in the extreme observations, it follows that the LAD parameter estimates are more resilient to outlying observations. (See Section A-1 for a brief discussion of the sample median.) In choosing the estimates, OLS squares each residual, and so the OLS estimates can be very sensitive to outlying observations, as we saw in Examples 9.8 and 9.10.

In addition to LAD being more computationally intensive than OLS, a second drawback of LAD is that all statistical inference involving the LAD estimators is justified only as the sample size grows. [The formulas are somewhat complicated and require matrix algebra, and we do not need them here. Koenker (2005) provides a comprehensive treatment.] Recall that, under the classical linear model assumptions, the OLS t statistics have exact t distributions, and F statistics have exact F distributions. While asymptotic versions of these statistics are available for LAD—and reported routinely by software packages that compute LAD estimates—these are justified only in large samples. Like the additional computational burden involved in computing LAD estimates, the lack of exact inference for LAD is only of minor concern, because most applications of LAD involve several hundred, if not

FIGURE 9.2 The OLS and LAD objective functions.



several thousand, observations. Of course, we might be pushing it if we apply large-sample approximations in an example such as Example 9.8, with $n = 32$. In a sense, this is not very different from OLS because, more often than not, we must appeal to large sample approximations to justify OLS inference whenever any of the CLM assumptions fail.

A more subtle but important drawback to LAD is that it does not always consistently estimate the parameters appearing in the conditional mean function, $\mathrm{E}(y|x_1, \ldots, x_k)$. As mentioned earlier, LAD is intended to estimate the effects on the conditional median. Generally, the mean and median are the same only when the distribution of y given the covariates $x_1, \ldots, x_k$ is symmetric about $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. (Equivalently, the population error term, u, is symmetric about zero.) Recall that OLS produces unbiased and consistent estimators of the parameters in the conditional mean whether or not the error distribution is symmetric; symmetry does not appear among the Gauss-Markov assumptions. When LAD and OLS are applied to cases with asymmetric distributions, the estimated partial effect of, say, x_1, obtained from LAD can be very different from the partial effect obtained from OLS. But such a difference could just reflect the difference between the median and the mean and might not have anything to do with outliers. See Computer Exercise C9 for an example.

If we assume that the population error u in model (9.2) is *independent* of $(x_1, \ldots, x_k)$, then the OLS and LAD slope estimates should differ only by sampling error whether or not the distribution of u is symmetric. The intercept estimates generally will be different to reflect the fact that, if the mean of u is zero, then its median is different from zero under asymmetry. Unfortunately, independence between the error and the explanatory variables is often unrealistically strong when LAD is applied. In particular, independence rules out heteroskedasticity, a problem that often arises in applications with asymmetric distributions.

An advantage that LAD has over OLS is that, because LAD estimates the median, it is easy to obtain partial effects—and predictions—using monotonic transformations. Here we consider the most common transformation, taking the natural log. Suppose that $\log(y)$ follows a linear model where the error has a zero conditional median:

$$\log(y) = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u \qquad [9.46]$$

$$\mathrm{Med}(u|\mathbf{x}) = 0, \qquad [9.47]$$

which implies that

$$\text{Med}[\log(y)|\mathbf{x}] = \beta_0 + \mathbf{x}\boldsymbol{\beta}.$$

A well-known feature of the conditional median—see, for example, Wooldridge (2010, Chapter 12)—is that it passes through increasing functions. Therefore,

$$\text{Med}(y|\mathbf{x}) = \exp(\beta_0 + \mathbf{x}\boldsymbol{\beta}). \qquad [9.48]$$

It follows that β_j is the semi-elasticity of $\text{Med}(y|\mathbf{x})$ with respect to x_j. In other words, the partial effect of x_j in the linear equation (9.46) can be used to uncover the partial effect in the nonlinear model (9.48). It is important to understand that this holds for any distribution of u such that (9.47) holds, and we need not assume u and $\mathbf{x}$ are independent. By contrast, if we specify a linear model for $\text{E}[\log(y)|\mathbf{x}]$ then, in general, there is no way to uncover $\text{E}(y|\mathbf{x})$. If we make a full distributional assumption for u given $\mathbf{x}$ then, in principle, we can recover $\text{E}(y|\mathbf{x})$. We covered the special case in equation (6.40) under the assumption that $\log(y)$ follows a classical linear model. However, in general there is no way to find $\text{E}(y|\mathbf{x})$ from a model for $\text{E}[\log(y)|\mathbf{x}]$, even though we can always obtain $\text{Med}(y|\mathbf{x})$ from $\text{Med}[\log(y)|\mathbf{x}]$. Problem 9 investigates how heteroskedasticity in a linear model for $\log(y)$ confounds our ability to find $\text{E}(y|\mathbf{x})$.

LAD is a special case of what is often called *robust regression*. Unfortunately, the way "robust" is used here can be confusing. In the statistics literature, a robust regression estimator is relatively insensitive to extreme observations. Effectively, observations with large residuals are given less weight than in least squares. [Berk (1990) contains an introductory treatment of estimators that are robust to outlying observations.] Based on our earlier discussion, in econometric parlance, LAD is not a robust estimator of the conditional mean because it requires extra assumptions in order to consistently estimate the conditional mean parameters. In Equation (9.2), either the distribution of u given $(x_1, \ldots, x_k)$ has to be symmetric about zero, or u must be independent of $(x_1, \ldots, x_k)$. Neither of these is required for OLS.

LAD is also a special case of *quantile regression*, which is used to estimate the effect of the x_j on different parts of the distribution—not just the median (or mean). For example, in a study to see how having access to a particular pension plan affects wealth, it could be that access affects high-wealth people differently from low-wealth people, and these effects both differ from the median person. Wooldridge (2010, Chapter 12) contains a treatment and examples of quantile regression.

Summary

We have further investigated some important specification and data issues that often arise in empirical cross-sectional analysis. Misspecified functional form makes the estimated equation difficult to interpret. Nevertheless, incorrect functional form can be detected by adding quadratics, computing RESET, or testing against a nonnested alternative model using the Davidson-MacKinnon test. No additional data collection is needed.

Solving the omitted variables problem is more difficult. In Section 9-2, we discussed a possible solution based on using a proxy variable for the omitted variable. Under reasonable assumptions, including the proxy variable in an OLS regression eliminates, or at least reduces, bias. The hurdle in applying this method is that proxy variables can be difficult to find. A general possibility is to use data on a dependent variable from a prior year.

Applied economists are often concerned with measurement error. Under the classical errors-in-variables (CEV) assumptions, measurement error in the dependent variable has no effect on the statistical properties of OLS. In contrast, under the CEV assumptions for an independent variable, the OLS estimator

for the coefficient on the mismeasured variable is biased toward zero. The bias in coefficients on the other variables can go either way and is difficult to determine.

Nonrandom samples from an underlying population can lead to biases in OLS. When sample selection is correlated with the error term u, OLS is generally biased and inconsistent. On the other hand, exogenous sample selection—which is either based on the explanatory variables or is otherwise independent of u—does not cause problems for OLS. Outliers in data sets can have large impacts on the OLS estimates, especially in small samples. It is important to at least informally identify outliers and to reestimate models with the suspected outliers excluded.

Least absolute deviations estimation is an alternative to OLS that is less sensitive to outliers and that delivers consistent estimates of conditional median parameters. In the past 20 years, with computational advances and improved understanding of the pros and cons of LAD and OLS, LAD is used more and more in empirical research–often as a supplement to OLS.

Key Terms

Attenuation Bias
Average Marginal Effect (AME)
Average Partial Effect (APE)
Classical Errors-in-Variables (CEV)
Complete Cases Estimator
Conditional Median
Davidson-MacKinnon Test
Endogenous Explanatory Variable
Endogenous Sample Selection
Exogenous Sample Selection
Functional Form Misspecification
Influential Observations
Lagged Dependent Variable
Least Absolute Deviations (LAD)
Measurement Error
Missing at Random (MAR)
Missing Completely at Random (MCAR)
Missing Data
Missing Indicator Method (MIM)
Multiplicative Measurement Error
Nonnested Models
Nonrandom Sample
Outliers
Plug-In Solution to the Omitted Variables Problem
Proxy Variable
Random Coefficient (Slope) Model
Regression Specification Error Test (RESET)
Stratified Sampling
Studentized Residuals

Problems

1 In Problem 11 in Chapter 4, the R-squared from estimating the model

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + \beta_3 profmarg + \beta_4 ceoten + \beta_5 comten + u,$$

using the data in CEOSAL2, was $R^2 = .353$ $(n = 177)$. When $ceoten^2$ and $comten^2$ are added, $R^2 = .375$. Is there evidence of functional form misspecification in this model?

2 Let us modify Computer Exercise C4 in Chapter 8 by using voting outcomes in 1990 for incumbents who were elected in 1988. Candidate A was elected in 1988 and was seeking reelection in 1990; *voteA90* is Candidate A's share of the two-party vote in 1990. The 1988 voting share of Candidate A is used as a proxy variable for quality of the candidate. All other variables are for the 1990 election. The following equations were estimated, using the data in VOTE2:

$$\widehat{voteA90} = 75.71 + .312\, prtystrA + 4.93\, democA$$
$$(9.25)\ (.046)\qquad (1.01)$$
$$-.929 \log(expendA) - 1.950 \log(expendB)$$
$$(.684)\qquad (.281)$$
$$n = 186,\ R^2 = .495,\ \overline{R}^2 = .483,$$

and

$$\widehat{voteA90} = 70.81 + .282\ prtystrA + 4.52\ democA$$
$$(10.01)\ \ (.052)\qquad\quad (1.06)$$
$$-.839\log(expendA) - 1.846\log(expendB) + .067\ voteA88$$
$$(.687)\qquad\quad (.292)\qquad\quad (.053)$$
$$n = 186,\ R^2 = .499,\ \overline{R}^2 = .485.$$

(i) Interpret the coefficient on *voteA88* and discuss its statistical significance.
(ii) Does adding *voteA88* have much effect on the other coefficients?

3 Let *math10* denote the percentage of students at a Michigan high school receiving a passing score on a standardized math test (see also Example 4.2). We are interested in estimating the effect of per-student spending on math performance. A simple model is

$$math10 = \beta_0 + \beta_1\log(expend) + \beta_2\log(enroll) + \beta_3 poverty + u,$$

where *poverty* is the percentage of students living in poverty.
(i) The variable *lnchprg* is the percentage of students eligible for the federally funded school lunch program. Why is this a sensible proxy variable for *poverty*?
(ii) The table that follows contains OLS estimates, with and without *lnchprg* as an explanatory variable.

Dependent Variable: *math10*		
Independent Variables	(1)	(2)
log(*expend*)	11.13 (3.30)	7.75 (3.04)
log(*enroll*)	.022 (.615)	–1.26 (.58)
lnchprg	—	–.324 (.036)
intercept	–69.24 (26.72)	–23.14 (24.99)
Observations *R*-squared	428 .0297	428 .1893

Explain why the effect of expenditures on *math10* is lower in column (2) than in column (1). Is the effect in column (2) still statistically greater than zero?
(iii) Does it appear that pass rates are lower at larger schools, other factors being equal? Explain.
(iv) Interpret the coefficient on *lnchprg* in column (2).
(v) What do you make of the substantial increase in R^2 from column (1) to column (2)?

4 The following equation explains weekly hours of television viewing by a child in terms of the child's age, mother's education, father's education, and number of siblings:

$$tvhours^* = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 motheduc + \beta_4 fatheduc + \beta_5 sibs + u.$$

We are worried that $tvhours^*$ is measured with error in our survey. Let *tvhours* denote the reported hours of television viewing per week.
(i) What do the classical errors-in-variables (CEV) assumptions require in this application?
(ii) Do you think the CEV assumptions are likely to hold? Explain.

5 In Example 4.4, we estimated a model relating number of campus crimes to student enrollment for a sample of colleges. The sample we used was not a random sample of colleges in the United States, because many schools in 1992 did not report campus crimes. Do you think that college failure to report crimes can be viewed as exogenous sample selection? Explain.

6 In the model (9.17), show that OLS consistently estimates α and β if a_i is uncorrelated with x_i and b_i is uncorrelated with x_i and x_i^2, which are weaker assumptions than (9.19). [*Hint*: Write the equation as in (9.18) and recall from Chapter 5 that sufficient for consistency of OLS for the intercept and slope is $\mathrm{E}(u_i) = 0$ and $\mathrm{Cov}(x_i, u_i) = 0$.]

7 Consider the simple regression model with classical measurement error, $y = \beta_0 + \beta_1 x^* + u$, where we have m measures on x^*. Write these as $z_h = x^* + e_h$, $h = 1, \ldots, m$. Assume that x^* is uncorrelated with $u, e_1, \ldots, e_m$, that the measurement errors are pairwise uncorrelated, and have the same variance, σ_e^2. Let $w = (z_1 + \cdots + z_m)/m$ be the average of the measures on x^*, so that, for each observation i, $w_i = (z_{i1} + \cdots + z_{im})/m$ is the average of the m measures. Let $\overline{\beta}_1$ be the OLS estimator from the simple regression y_i on 1, w_i, $i = 1, \ldots, n$, using a random sample of data.

(i) Show that

$$\mathrm{plim}(\overline{\beta}_1) = \beta_1 \left\{ \frac{\sigma_{x^*}^2}{[\sigma_{x^*}^2 + (\sigma_e^2/m)]} \right\}.$$

[*Hint:* The plim of $\overline{\beta}_1$ is $\mathrm{Cov}(w,y)/\mathrm{Var}(w)$.]

(ii) How does the inconsistency in $\overline{\beta}_1$ compare with that when only a single measure is available (that is, $m = 1$)? What happens as m grows? Comment.

8 The point of this exercise is to show that tests for functional form cannot be relied on as a general test for omitted variables. Suppose that, conditional on the explanatory variables x_1 and x_2, a linear model relating y to x_1 and x_2 satisfies the Gauss-Markov assumptions:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\ \mathrm{E}(u|x_1, x_2) &= 0 \\ \mathrm{Var}(u|x_1, x_2) &= \sigma^2. \end{aligned}$$

To make the question interesting, assume $\beta_2 \neq 0$.

Suppose further that x_2 has a simple linear relationship with x_1:

$$\begin{aligned} x_2 &= \delta_0 + \delta_1 x_1 + r \\ \mathrm{E}(r|x_1) &= 0 \\ \mathrm{Var}(r|x_1) &= \tau^2. \end{aligned}$$

(i) Show that

$$\mathrm{E}(y|x_1) = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1.$$

Under random sampling, what is the probability limit of the OLS estimator from the simple regression of y on x_1? Is the simple regression estimator generally consistent for β_1?

(ii) If you run the regression of y on x_1, x_1^2, what will be the probability limit of the OLS estimator of the coefficient on x_1^2? Explain.

(iii) Using substitution, show that we can write

$$y = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1 + u + \beta_2 r.$$

It can be shown that, if we define $v = u + \beta_2 r$ then $\mathrm{E}(v|x_1) = 0$, $\mathrm{Var}(v|x_1) = \sigma^2 + \beta_2^2\tau^2$. What consequences does this have for the t statistic on x_1^2 from the regression in part (ii)?

(iv) What do you conclude about adding a nonlinear function of x_1—in particular, x_1^2—in an attempt to detect omission of x_2?

9 Suppose that $\log(y)$ follows a linear model with a linear form of heteroskedasticity. We write this as

$$\log(y) = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u$$
$$u|\mathbf{x} \sim \text{Normal}[0,h(\mathbf{x})],$$

so that, conditional on $\mathbf{x}$, u has a normal distribution with mean (and median) zero but with variance $h(\mathbf{x})$ that depends on $\mathbf{x}$. Because $\text{Med}(u|\mathbf{x}) = 0$, equation (9.48) holds: $\text{Med}(y|\mathbf{x}) = \exp(\beta_0 + \mathbf{x}\boldsymbol{\beta})$. Further, using an extension of the result from Chapter 6, it can be shown that

$$\text{E}(y|\mathbf{x}) = \exp[\beta_0 + \mathbf{x}\boldsymbol{\beta} + h(\mathbf{x})/2].$$

(i) Given that $h(\mathbf{x})$ can be any positive function, is it possible to conclude $\partial\text{E}(y|\mathbf{x})/\partial x_j$ is the same sign as β_j?

(ii) Suppose $h(\mathbf{x}) = \delta_0 + \mathbf{x}\boldsymbol{\delta}$ (and ignore the problem that linear functions are not necessarily always positive). Show that a particular variable, say x_1, can have a negative effect on $\text{Med}(y|\mathbf{x})$ but a positive effect on $\text{E}(y|\mathbf{x})$.

(iii) Consider the case covered in Section 6-4, in which $h(\mathbf{x}) = \sigma^2$. How would you predict y using an estimate of $\text{E}(y|\mathbf{x})$? How would you predict y using an estimate of $\text{Med}(y|\mathbf{x})$? Which prediction is always larger?

10 This exercise shows that in a simple regression model, adding a dummy variable for missing data on the explanatory variable produces a consistent estimator of the slope coefficient if the "missingness" is unrelated to both the unobservable and observable factors affecting y. Let m be a variable such that $m = 1$ if we do not observe x and $m = 0$ if we observe x. We assume that y is always observed. The population model is

$$y = \beta_0 + \beta_1 x + u$$
$$\text{E}(u|x) = 0.$$

(i) Provide an interpretation of the stronger assumption

$$\text{E}(u|x,m) = 0.$$

In particular, what kind of missing data schemes would cause this assumption to fail?

(ii) Show that we can always write

$$y = \beta_0 + \beta_1(1 - m)x + \beta_1 mx + u.$$

(iii) Let (x_i, y_i, m_i): $i = 1, \ldots, n$ be random draws from the population, where x_i is missing when $m_i = 1$. Explain the nature of the variable $z_i = (1 - m_i)x_i$. In particular, what does this variable equal when x_i is missing?

(iv) Let $\rho = \text{P}(m = 1)$ and assume that m and x are independent. Show that

$$\text{Cov}[(1 - m)x,mx] = -\rho(1 - \rho)\mu_x,$$

where $\mu_x = \text{E}(x)$. What does this imply about estimating β_1 from the regression y_i on z_i, $i = 1, \ldots, n$?

(v) If m and x are independent, it can be shown that

$$mx = \delta_0 + \delta_1 m + v,$$

where v is uncorrelated with m and $z = (1 - m)x$. Explain why this makes m a suitable proxy variable for mx. What does this mean about the coefficient on z_i in the regression

$$y_i \text{ on } z_i, m_i, i = 1, \ldots, n?$$

(vi) Suppose for a population of children, y is a standardized test score, obtained from school records, and x is family income, which is reported voluntarily by families (and so some families do not report their income). Is it realistic to assume m and x are independent? Explain.

11 (i) In column (3) of Table 9.2, the coefficient on *educ* is .018 and it is statistically insignificant, and that on *IQ* is actually negative, −.0009, and also statistically insignificant. Explain what is happening.

(ii) What regression might you run that still includes an interaction to make the coefficients on *educ* and *IQ* more sensible? Explain.

Computer Exercises

C1 (i) Apply RESET from equation (9.3) to the model estimated in Computer Exercise C5 in Chapter 7. Is there evidence of functional form misspecification in the equation?

(ii) Compute a heteroskedasticity-robust form of RESET. Does your conclusion from part (i) change?

C2 Use the data set WAGE2 for this exercise.

(i) Use the variable *KWW* (the "knowledge of the world of work" test score) as a proxy for ability in place of *IQ* in Example 9.3. What is the estimated return to education in this case?

(ii) Now, use *IQ* and *KWW* together as proxy variables. What happens to the estimated return to education?

(iii) In part (ii), are *IQ* and *KWW* individually significant? Are they jointly significant?

C3 Use the data from JTRAIN for this exercise.

(i) Consider the simple regression model

$$\log(scrap) = \beta_0 + \beta_1 grant + u,$$

where *scrap* is the firm scrap rate and *grant* is a dummy variable indicating whether a firm received a job training grant. Can you think of some reasons why the unobserved factors in u might be correlated with *grant*?

(ii) Estimate the simple regression model using the data for 1988. (You should have 54 observations.) Does receiving a job training grant significantly lower a firm's scrap rate?

(iii) Now, add as an explanatory variable $\log(scrap_{87})$. How does this change the estimated effect of *grant*? Interpret the coefficient on *grant*. Is it statistically significant at the 5% level against the one-sided alternative H_1: $\beta_{grant} < 0$?

(iv) Test the null hypothesis that the parameter on $\log(scrap_{87})$ is one against the two-sided alternative. Report the p-value for the test.

(v) Repeat parts (iii) and (iv), using heteroskedasticity-robust standard errors, and briefly discuss any notable differences.

C4 Use the data for the year 1990 in INFMRT for this exercise.

(i) Reestimate equation (9.43), but now include a dummy variable for the observation on the District of Columbia (called *DC*). Interpret the coefficient on *DC* and comment on its size and significance.

(ii) Compare the estimates and standard errors from part (i) with those from equation (9.44). What do you conclude about including a dummy variable for a single observation?

C5 Use the data in RDCHEM to further examine the effects of outliers on OLS estimates and to see how LAD is less sensitive to outliers. The model is

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \beta_3 profmarg + u,$$

where you should first change *sales* to be in billions of dollars to make the estimates easier to interpret.

(i) Estimate the above equation by OLS, both with and without the firm having annual sales of almost $40 billion. Discuss any notable differences in the estimated coefficients.

(ii) Estimate the same equation by LAD, again with and without the largest firm. Discuss any important differences in estimated coefficients.

(iii) Based on your findings in parts (i) and (ii), would you say OLS or LAD is more resilient to outliers?

C6 Redo Example 4.10 by dropping schools where teacher benefits are less than 1% of salary.

(i) How many observations are lost?

(ii) Does dropping these observations have any important effects on the estimated tradeoff?

C7 Use the data in LOANAPP for this exercise.

(i) How many observations have $obrat > 40$, that is, other debt obligations more than 40% of total income?

(ii) Reestimate the model in part (iii) of Computer Exercise C8 in Chapter 7, excluding observations with $obrat > 40$. What happens to the estimate and t statistic on *white*?

(iii) Does it appear that the estimate of β_{white} is overly sensitive to the sample used?

C8 Use the data in TWOYEAR for this exercise.

(i) The variable *stotal* is a standardized test variable, which can act as a proxy variable for unobserved ability. Find the sample mean and standard deviation of *stotal*.

(ii) Run simple regressions of *jc* and *univ* on *stotal*. Are both college education variables statistically related to *stotal*? Explain.

(iii) Add *stotal* to equation (4.17) and test the hypothesis that the returns to two- and four-year colleges are the same against the alternative that the return to four-year colleges is greater. How do your findings compare with those from Section 4-4?

(iv) Add $stotal^2$ to the equation estimated in part (iii). Does a quadratic in the test score variable seem necessary?

(v) Add the interaction terms $stotal \cdot jc$ and $stotal \cdot univ$ to the equation from part (iii). Are these terms jointly significant?

(vi) What would be your final model that controls for ability through the use of *stotal*? Justify your answer.

C9 In this exercise, you are to compare OLS and LAD estimates of the effects of 401(k) plan eligibility on net financial assets. The model is

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 male + \beta_6 e401k + u.$$

(i) Use the data in 401KSUBS to estimate the equation by OLS and report the results in the usual form. Interpret the coefficient on *e401k*.

(ii) Use the OLS residuals to test for heteroskedasticity using the Breusch-Pagan test. Is u independent of the explanatory variables?

(iii) Estimate the equation by LAD and report the results in the same form as for OLS. Interpret the LAD estimate of β_6.

(iv) Reconcile your findings from parts (i) and (iii).

C10 You need to use two data sets for this exercise, JTRAIN2 and JTRAIN3. The former is the outcome of a job training experiment. The file JTRAIN3 contains observational data, where individuals themselves largely determine whether they participate in job training. The data sets cover the same time period.

(i) In the data set JTRAIN2, what fraction of the men received job training? What is the fraction in JTRAIN3? Why do you think there is such a big difference?

(ii) Using JTRAIN2, run a simple regression of *re78* on *train*. What is the estimated effect of participating in job training on real earnings?

(iii) Now add as controls to the regression in part (ii) the variables *re74*, *re75*, *educ*, *age*, *black*, and *hisp*. Does the estimated effect of job training on *re78* change much? How come? (*Hint:* Remember that these are experimental data.)

(iv) Do the regressions in parts (ii) and (iii) using the data in JTRAIN3, reporting only the estimated coefficients on *train*, along with their *t* statistics. What is the effect now of controlling for the extra factors, and why?

(v) Define $avgre = (re74 + re75)/2$. Find the sample averages, standard deviations, and minimum and maximum values in the two data sets. Are these data sets representative of the same populations in 1978?

(vi) Almost 96% of men in the data set JTRAIN2 have *avgre* less than \$10,000. Using only these men, run the regression

re78 on *train*, *re74*, *re75*, *educ*, *age*, *black*, *hisp*

and report the training estimate and its *t* statistic. Run the same regression for JTRAIN3, using only men with $avgre \leq 10$. For the subsample of low-income men, how do the estimated training effects compare across the experimental and nonexperimental data sets?

(vii) Now use each data set to run the simple regression *re78* on *train*, but only for men who were unemployed in 1974 and 1975. How do the training estimates compare now?

(viii) Using your findings from the previous regressions, discuss the potential importance of having comparable populations underlying comparisons of experimental and nonexperimental estimates.

C11 Use the data in MURDER only for the year 1993 for this question, although you will need to first obtain the lagged murder rate, say $mrdrte_{-1}$.

(i) Run the regression of *mrdrte* on *exec*, *unem*. What are the coefficient and *t* statistic on *exec*? Does this regression provide any evidence for a deterrent effect of capital punishment?

(ii) How many executions are reported for Texas during 1993? (Actually, this is the sum of executions for the current and past two years.) How does this compare with the other states? Add a dummy variable for Texas to the regression in part (i). Is its *t* statistic unusually large? From this, does it appear Texas is an "outlier"?

(iii) To the regression in part (i) add the lagged murder rate. What happens to $\hat{\beta}_{exec}$ and its statistical significance?

(iv) For the regression in part (iii), does it appear Texas is an outlier? What is the effect on $\hat{\beta}_{exec}$ from dropping Texas from the regression?

C12 Use the data in ELEM94_95 to answer this question. See also Computer Exercise C10 in Chapter 4.

(i) Using all of the data, run the regression *lavgsal* on *bs*, *lenrol*, *lstaff*, and *lunch*. Report the coefficient on *bs* along with its usual and heteroskedasticity-robust standard errors. What do you conclude about the economic and statistical significance of $\hat{\beta}_{bs}$?

(ii) Now drop the four observations with $bs > .5$, that is, where average benefits are (supposedly) more than 50% of average salary. What is the coefficient on *bs*? Is it statistically significant using the heteroskedasticity-robust standard error?

(iii) Verify that the four observations with $bs > .5$ are 68, 1,127, 1,508, and 1,670. Define four dummy variables for each of these observations. (You might call them *d*68, *d*1127, *d*1508, and *d*1670.) Add these to the regression from part (i) and verify that the OLS coefficients and standard errors on the other variables are identical to those in part (ii). Which of the four dummies has a *t* statistic statistically different from zero at the 5% level?

(iv) Verify that, in this data set, the data point with the largest studentized residual (largest *t* statistic on the dummy variable) in part (iii) has a large influence on the OLS estimates. (That is, run OLS using all observations except the one with the large studentized residual.) Does dropping, in turn, each of the other observations with $bs > .5$ have important effects?

(v) What do you conclude about the sensitivity of OLS to a single observation, even with a large sample size?

(vi) Verify that the LAD estimator is not sensitive to the inclusion of the observation identified in part (iii).

C13 Use the data in CEOSAL2 to answer this question.

(i) Estimate the model

$$lsalary = \beta_0 + \beta_1 lsales + \beta_2 lmktval + \beta_3 ceoten + \beta_4 ceoten^2 + u$$

by OLS using all of the observations, where *lsalary*, *lsales*, and *lmktval* are all natural logarithms. Report the results in the usual form with the usual OLS standard errors. (You may verify that the heteroskedasticity-robust standard errors are similar.)

(ii) In the regression from part (i) obtain the studentized residuals; call these str_i. How many studentized residuals are above 1.96 in absolute value? If the studentized residuals were independent draws from a standard normal distribution, about how many would you expect to be above two in absolute value with 177 draws?

(iii) Reestimate the equation in part (i) by OLS using only the observations with $|str_i| \leq 1.96$. How do the coefficients compare with those in part (i)?

(iv) Estimate the equation in part (i) by LAD, using all of the data. Is the estimate of β_1 closer to the OLS estimate using the full sample or the restricted sample? What about for β_3?

(v) Evaluate the following statement: "Dropping outliers based on extreme values of studentized residuals makes the resulting OLS estimates closer to the LAD estimates on the full sample."

C14 Use the data in ECONMATH to answer this question. The population model is

$$score = \beta_0 + \beta_1 act + u.$$

(i) For how many students is the ACT score missing? What is the fraction of the sample? Define a new variable, *actmiss*, which equals one if *act* is missing, and zero otherwise.

(ii) Create a new variable, say *act*0, which is the *act* score when *act* is reported and zero when *act* is missing. Find the average of *act*0 and compare it with the average for *act*.

(iii) Run the simple regression of *score* on *act* using only the complete cases. What do you obtain for the slope coefficient and its heteroskedasticity-robust standard error?

(iv) Run the simple regression of *score* on *act*0 using all of the cases. Compare the slope coefficient with that in part (iii) and comment.

(v) Now use all of the cases and run the regression

$$score_i \text{ on } act0_i, actmiss_i.$$

What is the slope estimate on $act0_i$? How does it compare with the answers in parts (iii) and (iv)?

(vi) Comparing regressions in parts (iii) and (v), does using all cases and adding the missing data estimator improve estimation of β_1?

(vii) If you add the variable *colgpa* to the regressions in parts (iii) and (v), does this change your answer to part (vi)?

PART 2

Regression Analysis with Time Series Data

Now that we have a solid understanding of how to use the multiple regression model for cross-sectional applications, we can turn to the econometric analysis of time series data. Because we will rely heavily on the method of ordinary least squares, most of the work concerning mechanics and inference has already been done. However, as we noted in Chapter 1, time series data have certain characteristics that cross-sectional data do not, and these can require special attention when applying OLS.

Chapter 10 covers basic regression analysis and gives attention to problems unique to time series data. We provide a set of Gauss-Markov and classical linear model assumptions for time series applications. The problems of functional form, dummy variables, trends, and seasonality are also discussed.

Because certain time series models necessarily violate the Gauss-Markov assumptions, Chapter 11 describes the nature of these violations and presents the large sample properties of ordinary least squares. As we can no longer assume random sampling, we must cover conditions that restrict the temporal correlation in a time series in order to ensure that the usual asymptotic analysis is valid.

Chapter 12 turns to an important new problem: serial correlation in the error terms in time series regressions. We discuss the consequences, ways of testing, and methods for dealing with serial correlation. Chapter 12 also contains an explanation of how heteroskedasticity can arise in time series models.

CHAPTER 10

Basic Regression Analysis with Time Series Data

In this chapter, we begin to study the properties of OLS for estimating linear regression models using time series data. In Section 10-1, we discuss some conceptual differences between time series and cross-sectional data. Section 10-2 provides some examples of time series regressions that are often estimated in the empirical social sciences. We then turn our attention to the finite sample properties of the OLS estimators and state the Gauss-Markov assumptions and the classical linear model assumptions for time series regression. Although these assumptions have features in common with those for the cross-sectional case, they also have some significant differences that we will need to highlight.

In addition, we return to some issues that we treated in regression with cross-sectional data, such as how to use and interpret the logarithmic functional form and dummy variables. The important topics of how to incorporate trends and account for seasonality in multiple regression are taken up in Section 10-5.

10-1 The Nature of Time Series Data

An obvious characteristic of time series data that distinguishes them from cross-sectional data is temporal ordering. For example, in Chapter 1, we briefly discussed a time series data set on employment, the minimum wage, and other economic variables for Puerto Rico. In this data set, we must know that the data for 1970 immediately precede the data for 1971. For analyzing time series data in the social sciences, we must recognize that the past can affect the future, but not vice versa (unlike in the *Star Trek* universe). To emphasize the proper ordering of time series data, Table 10.1 gives a partial listing

TABLE 10.1 Partial Listing of Data on U.S. Inflation and Unemployment Rates, 1948–2017

Year	Inflation	Unemployment
1948	8.1	3.8
1949	−1.2	5.9
1950	1.3	5.3
1951	7.9	3.3
.	.	.
.	.	.
.	.	.
2012	2.1	8.1
2013	1.5	7.4
2014	1.6	6.2
2015	0.1	5.3
2016	1.3	4.9
2017	2.1	4.4

of the data on U.S. inflation and unemployment rates from various editions of the *Economic Report of the President*, including the 2018 *Report* (Tables B-10 and B-11).

Another difference between cross-sectional and time series data is more subtle. In Chapters 3 and 4, we studied statistical properties of the OLS estimators based on the notion that samples were randomly drawn from the appropriate population. Understanding why cross-sectional data should be viewed as random outcomes is fairly straightforward: a different sample drawn from the population will generally yield different values of the independent and dependent variables (such as education, experience, wage, and so on). Therefore, the OLS estimates computed from different random samples will generally differ, and this is why we consider the OLS estimators to be random variables.

How should we think about randomness in time series data? Certainly, economic time series satisfy the intuitive requirements for being outcomes of random variables. For example, today we do not know what the Dow Jones Industrial Average will be at the close of the next trading day. We do not know what the annual growth in output will be in Canada during the coming year. Because the outcomes of these variables are not foreknown, they should clearly be viewed as random variables.

Formally, a sequence of random variables indexed by time is called a **stochastic process** or a **time series process**. ("Stochastic" is a synonym for random.) When we collect a time series data set, we obtain one possible outcome, or *realization*, of the stochastic process. We can only see a single realization because we cannot go back in time and start the process over again. (This is analogous to cross-sectional analysis where we can collect only one random sample.) However, if certain conditions in history had been different, we would generally obtain a different realization for the stochastic process, and this is why we think of time series data as the outcome of random variables. The set of all possible realizations of a time series process plays the role of the population in cross-sectional analysis. The sample size for a time series data set is the number of time periods over which we observe the variables of interest.

10-2 Examples of Time Series Regression Models

In this section, we discuss two examples of time series models that have been useful in empirical time series analysis and that are easily estimated by ordinary least squares. We will study additional models in Chapter 11.

10-2a Static Models

Suppose that we have time series data available on two variables, say y and z, where y_t and z_t are dated contemporaneously. A **static model** relating y to z is

$$y_t = \beta_0 + \beta_1 z_t + u_t, t = 1, 2, \ldots, n. \quad [10.1]$$

The name "static model" comes from the fact that we are modeling a contemporaneous relationship between y and z. Usually, a static model is postulated when a change in z at time t is believed to have an immediate effect on y: $\Delta y_t = \beta_1 \Delta z_t$, when $\Delta u_t = 0$. Static regression models are also used when we are interested in knowing the tradeoff between y and z.

An example of a static model is the *static Phillips curve*, given by

$$inf_t = \beta_0 + \beta_1 unem_t + u_t, \quad [10.2]$$

where inf_t is the annual inflation rate and $unem_t$ is the annual unemployment rate. This form of the Phillips curve assumes a constant *natural rate of unemployment* and constant inflationary expectations, and it can be used to study the contemporaneous tradeoff between inflation and unemployment. [See, for example, Mankiw (1994, Section 11-2).]

Naturally, we can have several explanatory variables in a static regression model. Let $mrdrte_t$ denote the murders per 10,000 people in a particular city during year t, let $convrte_t$ denote the murder conviction rate, let $unem_t$ be the local unemployment rate, and let $yngmle_t$ be the fraction of the population consisting of males between the ages of 18 and 25. Then, a static multiple regression model explaining murder rates is

$$mrdrte_t = \beta_0 + \beta_1 convrte_t + \beta_2 unem_t + \beta_3 yngmle_t + u_t. \quad [10.3]$$

Using a model such as this, we can hope to estimate, for example, the ceteris paribus effect of an increase in the conviction rate on a particular criminal activity.

10-2b Finite Distributed Lag Models

In a **finite distributed lag (FDL) model**, we allow one or more variables to affect y with a lag. For example, for annual observations, consider the model

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + u_t, \quad [10.4]$$

where gfr_t is the general fertility rate (children born per 1,000 women of childbearing age) and pe_t is the real dollar value of the personal tax exemption. The idea is to see whether, in the aggregate, the decision to have children is linked to the tax value of having a child. Equation (10.4) recognizes that, for both biological and behavioral reasons, decisions to have children would not immediately result from changes in the personal exemption.

Equation (10.4) is an example of the model

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t, \quad [10.5]$$

which is an FDL *of order two*. To interpret the coefficients in (10.5), suppose that z is a constant, equal to c, in all time periods before time t. At time t, z increases by one unit to $c + 1$ and then reverts to its previous level at time $t + 1$. (That is, the increase in z is temporary.) More precisely,

$$\ldots, z_{t-2} = c, z_{t-1} = c, z_t = c + 1, z_{t+1} = c, z_{t+2} = c, \ldots.$$

To focus on the ceteris paribus effect of z on y, we set the error term in each time period to zero. Then,

$$\begin{aligned} y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\ y_t &= \alpha_0 + \delta_0(c+1) + \delta_1 c + \delta_2 c, \\ y_{t+1} &= \alpha_0 + \delta_0 c + \delta_1(c+1) + \delta_2 c, \\ y_{t+2} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2(c+1), \\ y_{t+3} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \end{aligned}$$

and so on. From the first two equations, $y_t - y_{t-1} = \delta_0$, which shows that δ_0 is the immediate change in y due to the one-unit increase in z at time t. Usually, δ_0 is called the **impact propensity** or **impact multiplier**.

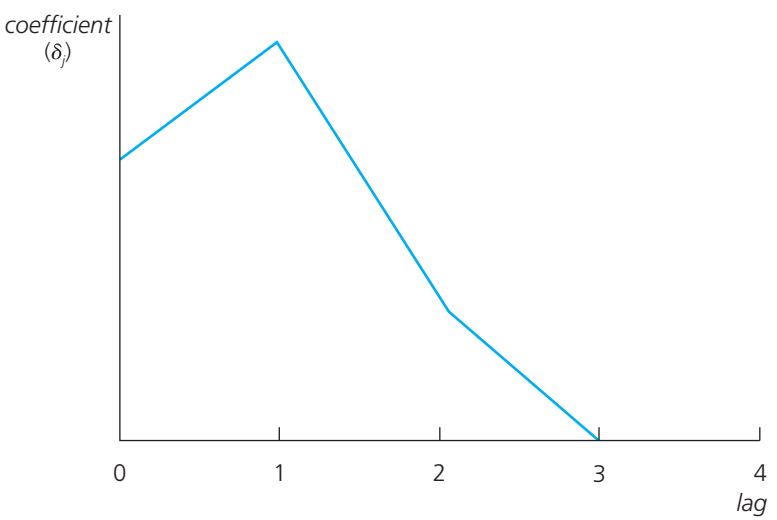
Similarly, $\delta_1 = y_{t+1} - y_{t-1}$ is the change in y one period after the temporary change and $\delta_2 = y_{t+2} - y_{t-1}$ is the change in y two periods after the change. At time $t + 3$, y has reverted back to its initial level: $y_{t+3} = y_{t-1}$. This is because we have assumed that only two lags of z appear in (10.5). When we graph the δ_j as a function of j, we obtain the **lag distribution**, which summarizes the dynamic effect that a temporary increase in z has on y. A possible lag distribution for the FDL of order two is given in Figure 10.1. (Of course, we would never know the parameters δ_j; instead, we will estimate the δ_j and then plot the estimated lag distribution.)

The lag distribution in Figure 10.1 implies that the largest effect is at the first lag. The lag distribution has a useful interpretation. If we standardize the initial value of y at $y_{t-1} = 0$, the lag distribution traces out all subsequent values of y due to a one-unit, temporary increase in z.

We are also interested in the change in y due to a *permanent* increase in z. Before time t, z equals the constant c. At time t, z increases permanently to $c + 1$: $z_s = c$, $s < t$ and $z_s = c + 1$, $s \geq t$. Again, setting the errors to zero, we have

$$\begin{aligned} y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\ y_t &= \alpha_0 + \delta_0(c+1) + \delta_1 c + \delta_2 c, \\ y_{t+1} &= \alpha_0 + \delta_0(c+1) + \delta_1(c+1) + \delta_2 c, \\ y_{t+2} &= \alpha_0 + \delta_0(c+1) + \delta_1(c+1) + \delta_2(c+1), \end{aligned}$$

FIGURE 10.1 A lag distribution with two nonzero lags. The maximum effect is at the first lag.



and so on. With the permanent increase in z, after one period, y has increased by $\delta_0 + \delta_1$, and after two periods, y has increased by $\delta_0 + \delta_1 + \delta_2$. There are no further changes in y after two periods. This shows that the sum of the coefficients on current and lagged z, $\delta_0 + \delta_1 + \delta_2$, is the *long-run* change in y given a permanent increase in z and is called the **long-run propensity (LRP)** or **long-run multiplier**. The LRP is often of interest in distributed lag models.

As an example, in equation (10.4), δ_0 measures the immediate change in fertility due to a one-dollar increase in *pe*. As we mentioned earlier, there are reasons to believe that δ_0 is small, if not zero. But δ_1 or δ_2, or both, might be positive. If *pe permanently* increases by one dollar, then, after two years, *gfr* will have changed by $\delta_0 + \delta_1 + \delta_2$. This model assumes that there are no further changes after two years. Whether this is actually the case is an empirical matter.

An FDL of order q is written as

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \cdots + \delta_q z_{t-q} + u_t. \qquad [10.6]$$

This contains the static model as a special case by setting $\delta_1, \delta_2, \ldots, \delta_q$ equal to zero. Sometimes, a primary purpose for estimating a distributed lag model is to test whether z has a lagged effect on y. The impact propensity is always the coefficient on the contemporaneous z, δ_0. Occasionally, we omit z_t from (10.6), in which case the impact propensity is zero. In the general case, the lag distribution can be plotted by graphing the (estimated) δ_j as a function of j. For any horizon h, we can define the **cumulative effect** as $\delta_0 + \delta_1 + \cdots + \delta_h$, which is interpreted as the change in the expected outcome h periods after a permanent, one-unit increase in x. Once the δ_j have been estimated, one may plot the estimated cumulative effects as a function of h. The LRP is the cumulative effect after all changes have taken place; it is simply the sum of all of the coefficients on the z_{t-j}:

$$\text{LRP} = \delta_0 + \delta_1 + \cdots + \delta_q. \qquad [10.7]$$

GOING FURTHER 10.1

In an equation for annual data, suppose that

$$int_t = 1.6 + .48\, inf_t - .15\, inf_{t-1} + .32\, inf_{t-2} + u_t,$$

where *int* is an interest rate and *inf* is the inflation rate. What are the impact and long-run propensities?

Because of the often substantial correlation in z at different lags—that is, due to multicollinearity in (10.6)—it can be difficult to obtain precise estimates of the individual δ_j. Interestingly, even when the δ_j cannot be precisely estimated, we can often get good estimates of the LRP. We will see an example later.

We can have more than one explanatory variable appearing with lags, or we can add contemporaneous variables to an FDL model. For example, the average education level for women of childbearing age could be added to (10.4), which allows us to account for changing education levels for women.

10-2c A Convention about the Time Index

When models have lagged explanatory variables (and, as we will see in Chapter 11, for models with lagged y), confusion can arise concerning the treatment of initial observations. For example, if in (10.5) we assume that the equation holds starting at $t = 1$, then the explanatory variables for the first time period are z_1, z_0, and z_{-1}. Our convention will be that these are the initial values in our sample, so that we can always start the time index at $t = 1$. In practice, this is not very important because regression packages automatically keep track of the observations available for estimating models with lags. But for this and the next two chapters, we need some convention concerning the first time period being represented by the regression equation.

10-3 Finite Sample Properties of OLS under Classical Assumptions

In this section, we give a complete listing of the finite sample, or small sample, properties of OLS under standard assumptions. We pay particular attention to how the assumptions must be altered from our cross-sectional analysis to cover time series regressions.

10-3a Unbiasedness of OLS

The first assumption simply states that the time series process follows a model that is linear in its parameters.

Assumption TS.1 **Linear in Parameters**

The stochastic process $\{(x_{t1}, x_{t2}, \ldots, x_{tk}, y_t): t = 1, 2, \ldots, n\}$ follows the linear model

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t, \qquad [10.8]$$

where $\{u_t: t = 1, 2, \ldots, n\}$ is the sequence of errors or disturbances. Here, n is the number of observations (time periods).

In the notation x_{tj}, t denotes the time period, and j is, as usual, a label to indicate one of the k explanatory variables. The terminology used in cross-sectional regression applies here: y_t is the dependent variable, explained variable, or regressand; the x_{tj} are the independent variables, explanatory variables, or regressors.

We should think of Assumption TS.1 as being essentially the same as Assumption MLR.1 (the first cross-sectional assumption), but we are now specifying a linear model for time series data. The examples covered in Section 10-2 can be cast in the form of (10.8) by appropriately defining x_{tj}. For example, equation (10.5) is obtained by setting $x_{t1} = z_t$, $x_{t2} = z_{t-1}$, and $x_{t3} = z_{t-2}$.

To state and discuss several of the remaining assumptions, we let $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tk})$ denote the set of all independent variables in the equation at time t. Further, $\mathbf{X}$ denotes the collection of all independent variables for all time periods. It is useful to think of $\mathbf{X}$ as being an array, with n rows and k columns. This reflects how time series data are stored in econometric software packages: the t^{th} row of $\mathbf{X}$ is $\mathbf{x}_t$, consisting of all independent variables for time period t. Therefore, the first row of $\mathbf{X}$ corresponds to $t = 1$, the second row to $t = 2$, and the last row to $t = n$. An example is given in Table 10.2, using $n = 8$ and the explanatory variables in equation (10.3).

TABLE 10.2 Example of X for the Explanatory Variables in Equation (10.3)

t	*convrte*	*unem*	*yngmle*
1	.46	.074	.12
2	.42	.071	.12
3	.42	.063	.11
4	.47	.062	.09
5	.48	.060	.10
6	.50	.059	.11
7	.55	.058	.12
8	.56	.059	.13

Naturally, as with cross-sectional regression, we need to rule out perfect collinearity among the regressors.

Assumption TS.2 **No Perfect Collinearity**

In the sample (and therefore in the underlying time series process), no independent variable is constant nor a perfect linear combination of the others.

We discussed this assumption at length in the context of cross-sectional data in Chapter 3. The issues are essentially the same with time series data. Remember, Assumption TS.2 does allow the explanatory variables to be correlated, but it rules out *perfect* correlation in the sample.

The final assumption for unbiasedness of OLS is the time series analog of Assumption MLR.4, and it also obviates the need for random sampling in Assumption MLR.2.

Assumption TS.3 **Zero Conditional Mean**

For each t, the expected value of the error u_t, given the explanatory variables for *all* time periods, is zero. Mathematically,

$$\mathrm{E}(u_t|\mathbf{X}) = 0, t = 1, 2, \ldots, n. \qquad [10.9]$$

This is a crucial assumption, and we need to have an intuitive grasp of its meaning. As in the cross-sectional case, it is easiest to view this assumption in terms of uncorrelatedness: Assumption TS.3 implies that the error at time t, u_t, is uncorrelated with each explanatory variable in *every* time period. The fact that this is stated in terms of the conditional expectation means that we must also correctly specify the functional relationship between y_t and the explanatory variables. If u_t is independent of $\mathbf{X}$ and $\mathrm{E}(u_t) = 0$, then Assumption TS.3 automatically holds.

Given the cross-sectional analysis from Chapter 3, it is not surprising that we require u_t to be uncorrelated with the explanatory variables also dated at time t: in conditional mean terms,

$$\mathrm{E}(u_t|x_{t1}, \ldots, x_{tk}) = \mathrm{E}(u_t|\mathbf{x}_t) = 0. \qquad [10.10]$$

When (10.10) holds, we say that the x_{tj} are **contemporaneously exogenous**. Equation (10.10) implies that u_t and the explanatory variables are contemporaneously uncorrelated: $\mathrm{Corr}(x_{tj},u_t) = 0$, for all j.

Assumption TS.3 requires more than contemporaneous exogeneity: u_t must be uncorrelated with x_{sj}, even when $s \neq t$. This is a strong sense in which the explanatory variables must be exogenous, and when TS.3 holds, we say that the explanatory variables are **strictly exogenous**. In Chapter 11, we will demonstrate that (10.10) is sufficient for proving consistency of the OLS estimator. But to show that OLS is unbiased, we need the strict exogeneity assumption.

In the cross-sectional case, we did not explicitly state how the error term for, say, person i, u_i, is related to the explanatory variables for *other* people in the sample. This was unnecessary because with random sampling (Assumption MLR.2), u_i is *automatically* independent of the explanatory variables for observations other than i. In a time series context, random sampling is almost never appropriate, so we must explicitly assume that the expected value of u_t is not related to the explanatory variables in any time periods.

It is important to see that Assumption TS.3 puts no restriction on correlation in the independent variables or in the u_t across time. Assumption TS.3 only says that the average value of u_t is unrelated to the independent variables in all time periods.

Anything that causes the unobservables at time t to be correlated with any of the explanatory variables in any time period causes Assumption TS.3 to fail. Two leading candidates for failure are omitted variables and measurement error in some of the regressors. But the strict exogeneity assumption can also fail for other, less obvious reasons. In the simple static regression model

$$y_t = \beta_0 + \beta_1 z_t + u_t,$$

Assumption TS.3 requires not only that u_t and z_t are uncorrelated, but that u_t is also uncorrelated with past and future values of z. This has two implications. First, z can have no lagged effect on y. If z does have a lagged effect on y, then we should estimate a distributed lag model. A more subtle point is that strict exogeneity excludes the possibility that changes in the error term today can cause future changes in z. This effectively rules out feedback from y to future values of z. For example, consider a simple static model to explain a city's murder rate in terms of police officers per capita:

$$mrdrte_t = \beta_0 + \beta_1 polpc_t + u_t.$$

It may be reasonable to assume that u_t is uncorrelated with $polpc_t$ and even with past values of $polpc_t$; for the sake of argument, assume this is the case. But suppose that the city adjusts the size of its police force based on past values of the murder rate. This means that, say, $polpc_{t+1}$ might be correlated with u_t (because a higher u_t leads to a higher $mrdrte_t$). If this is the case, Assumption TS.3 is generally violated.

There are similar considerations in distributed lag models. Usually, we do not worry that u_t might be correlated with past z because we are controlling for past z in the model. But feedback from u to future z is always an issue.

Explanatory variables that are strictly exogenous cannot react to what has happened to y in the past. A factor such as the amount of rainfall in an agricultural production function satisfies this requirement: rainfall in any future year is not influenced by the output during the current or past years. But something like the amount of labor input might not be strictly exogenous, as it is chosen by the farmer, and the farmer may adjust the amount of labor based on last year's yield. Policy variables, such as growth in the money supply, expenditures on welfare, and highway speed limits, are often influenced by what has happened to the outcome variable in the past. In the social sciences, many explanatory variables may very well violate the strict exogeneity assumption.

Even though Assumption TS.3 can be unrealistic, we begin with it in order to conclude that the OLS estimators are unbiased. Most treatments of static and FDL models assume TS.3 by making the stronger assumption that the explanatory variables are nonrandom, or fixed in repeated samples. The nonrandomness assumption is obviously false for time series observations; Assumption TS.3 has the advantage of being more realistic about the random nature of the x_{tj}, while it isolates the necessary assumption about how u_t and the explanatory variables are related in order for OLS to be unbiased.

THEOREM 10.1

UNBIASEDNESS OF OLS

Under Assumptions TS.1, TS.2, and TS.3, the OLS estimators are unbiased conditional on $\mathbf{X}$, and therefore unconditionally as well when the expectations exist: $\mathrm{E}(\hat{\beta}_j) = \beta_j, j = 0, 1, \ldots, k$.

GOING FURTHER 10.2

In the FDL model $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + u_t$, what do we need to assume about the sequence $\{z_0, z_1, \ldots, z_n\}$ in order for Assumption TS.3 to hold?

The proof of this theorem is essentially the same as that for Theorem 3.1 in Chapter 3, and so we omit it. When comparing Theorem 10.1 to Theorem 3.1, we have been able to drop the random sampling assumption by assuming that, for each t, u_t has zero mean given the explanatory variables at all time periods. If this assumption does not hold, OLS cannot be shown to be unbiased.

The analysis of omitted variables bias, which we covered in Section 3-3, is essentially the same in the time series case. In particular, Table 3.2 and the discussion surrounding it can be used as before to determine the directions of bias due to omitted variables.

10-3b The Variances of the OLS Estimators and the Gauss-Markov Theorem

We need to add two assumptions to round out the Gauss-Markov assumptions for time series regressions. The first one is familiar from cross-sectional analysis.

Assumption TS.4 Homoskedasticity

Conditional on $\mathbf{X}$, the variance of u_t is the same for all t: $\text{Var}(u_t|\mathbf{X}) = \text{Var}(u_t) = \sigma^2$, $t = 1, 2, \ldots, n$.

This assumption means that $\text{Var}(u_t|\mathbf{X})$ cannot depend on $\mathbf{X}$—it is sufficient that u_t and $\mathbf{X}$ are independent—*and* that $\text{Var}(u_t)$ is constant over time. When TS.4 does not hold, we say that the errors are *heteroskedastic*, just as in the cross-sectional case. For example, consider an equation for determining three-month T-bill rates $(i3_t)$ based on the inflation rate (inf_t) and the federal deficit as a percentage of gross domestic product (def_t):

$$i3_t = \beta_0 + \beta_1 inf_t + \beta_2 def_t + u_t. \qquad [10.11]$$

Among other things, Assumption TS.4 requires that the unobservables affecting interest rates have a constant variance over time. Because policy regime changes are known to affect the variability of interest rates, this assumption might very well be false. Further, it could be that the variability in interest rates depends on the level of inflation or relative size of the deficit. This would also violate the homoskedasticity assumption.

When $\text{Var}(u_t|\mathbf{X})$ does depend on $\mathbf{X}$, it often depends on the explanatory variables at time t, $\mathbf{x}_t$. In Chapter 12, we will see that the tests for heteroskedasticity from Chapter 8 can also be used for time series regressions, at least under certain assumptions.

The final Gauss-Markov assumption for time series analysis is new.

Assumption TS.5 No Serial Correlation

Conditional on $\mathbf{X}$, the errors in two different time periods are uncorrelated: $\text{Corr}(u_t,u_s|\mathbf{X}) = 0$, for all $t \neq s$.

The easiest way to think of this assumption is to ignore the conditioning on $\mathbf{X}$. Then, Assumption TS.5 is simply

$$\text{Corr}(u_t,u_s) = 0, \text{ for all } t \neq s. \qquad [10.12]$$

(This is how the no serial correlation assumption is stated when $\mathbf{X}$ is treated as nonrandom.) When considering whether Assumption TS.5 is likely to hold, we focus on equation (10.12) because of its simple interpretation.

When (10.12) is false, we say that the errors in (10.8) suffer from **serial correlation**, or **autocorrelation**, because they are correlated across time. Consider the case of errors from adjacent time periods. Suppose that when $u_{t-1} > 0$ then, on average, the error in the next time period, u_t, is also positive. Then, $\text{Corr}(u_t,u_{t-1}) > 0$, and the errors suffer from serial correlation. In equation (10.11), this means that if interest rates are unexpectedly high for this period, then they are likely to be above

average (for the given levels of inflation and deficits) for the next period. This turns out to be a reasonable characterization for the error terms in many time series applications, which we will see in Chapter 12. For now, we assume TS.5.

Importantly, Assumption TS.5 assumes nothing about temporal correlation in the *independent* variables. For example, in equation (10.11), inf_t is almost certainly correlated across time. But this has nothing to do with whether TS.5 holds.

A natural question that arises is: in Chapters 3 and 4, why did we not assume that the errors for different cross-sectional observations are uncorrelated? The answer comes from the random sampling assumption: under random sampling, u_i and u_h are independent for any two observations i and h. It can also be shown that, under random sampling, the errors for different observations are independent conditional on the explanatory variables in the sample. Thus, for our purposes, we consider serial correlation only to be a potential problem for regressions with time series data. (In Chapters 13 and 14, the serial correlation issue will come up in connection with panel data analysis.)

Assumptions TS.1 through TS.5 are the appropriate Gauss-Markov assumptions for time series applications, but they have other uses as well. Sometimes, TS.1 through TS.5 are satisfied in cross-sectional applications, even when random sampling is not a reasonable assumption, such as when the cross-sectional units are large relative to the population. Suppose that we have a cross-sectional data set at the city level. It might be that correlation exists across cities within the same state in some of the explanatory variables, such as property tax rates or per capita welfare payments. Correlation of the explanatory variables across observations does not cause problems for verifying the Gauss-Markov assumptions, provided the error terms are uncorrelated across cities. However, in this chapter, we are primarily interested in applying the Gauss-Markov assumptions to time series regression problems.

THEOREM 10.2

OLS SAMPLING VARIANCES

Under the time series Gauss-Markov Assumptions TS.1 through TS.5, the variance of $\hat{\beta}_j$, conditional on $\mathbf{X}$, is

$$\text{Var}(\hat{\beta}_j|\mathbf{X}) = \sigma^2/[\text{SST}_j(1 - R_j^2)], j = 1, \ldots, k, \qquad [10.13]$$

where SST_j is the total sum of squares of x_{ij} and R_j^2 is the R-squared from the regression of x_j on the other independent variables.

Equation (10.13) is the same variance we derived in Chapter 3 under the cross-sectional Gauss-Markov assumptions. Because the proof is very similar to the one for Theorem 3.2, we omit it. The discussion from Chapter 3 about the factors causing large variances, including multicollinearity among the explanatory variables, applies immediately to the time series case.

The usual estimator of the error variance is also unbiased under Assumptions TS.1 through TS.5, and the Gauss-Markov Theorem holds.

THEOREM 10.3

UNBIASED ESTIMATION OF σ^2

Under Assumptions TS.1 through TS.5, the estimator $\hat{\sigma}^2 = \text{SSR}/df$ is an unbiased estimator of σ^2, where $df = n - k - 1$.

THEOREM 10.4

GAUSS-MARKOV THEOREM

Under Assumptions TS.1 through TS.5, the OLS estimators are the best linear unbiased estimators conditional on $\mathbf{X}$.

GOING FURTHER 10.3

In the FDL model $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + u_t$, explain the nature of any multicollinearity in the explanatory variables.

The bottom line here is that OLS has the same desirable finite sample properties under TS.1 through TS.5 that it has under MLR.1 through MLR.5.

10-3c Inference under the Classical Linear Model Assumptions

In order to use the usual OLS standard errors, t statistics, and F statistics, we need to add a final assumption that is analogous to the normality assumption we used for cross-sectional analysis.

Assumption TS.6 Normality

The errors u_t are independent of $\mathbf{X}$ and are independently and identically distributed as Normal$(0,\sigma^2)$.

Assumption TS.6 implies TS.3, TS.4, and TS.5, but it is stronger because of the independence and normality assumptions.

THEOREM 10.5 NORMAL SAMPLING DISTRIBUTIONS

Under Assumptions TS.1 through TS.6, the CLM assumptions for time series, the OLS estimators are normally distributed, conditional on $\mathbf{X}$. Further, under the null hypothesis, each t statistic has a t distribution, and each F statistic has an F distribution. The usual construction of confidence intervals is also valid.

The implications of Theorem 10.5 are of utmost importance. It implies that, when Assumptions TS.1 through TS.6 hold, everything we have learned about estimation and inference for cross-sectional regressions applies directly to time series regressions. Thus, t statistics can be used for testing statistical significance of individual explanatory variables, and F statistics can be used to test for joint significance.

Just as in the cross-sectional case, the usual inference procedures are only as good as the underlying assumptions. The classical linear model assumptions for time series data are much more restrictive than those for cross-sectional data—in particular, the strict exogeneity and no serial correlation assumptions can be unrealistic. Nevertheless, the CLM framework is a good starting point for many applications.

EXAMPLE 10.1 Static Phillips Curve

To determine whether there is a tradeoff, on average, between unemployment and inflation, we can test H_0: $\beta_1 = 0$ against H_1: $\beta_1 < 0$ in equation (10.2). If the classical linear model assumptions hold, we can use the usual OLS t statistic.

We use the file PHILLIPS to estimate equation (10.2), restricting ourselves to the data through 2006. (In later exercises, for example, Computer Exercise C12 and Computer Exercise C10 in Chapter 11 you are asked to use all years through 2017. In Chapter 18, we use the years 2007 through 2017 in various forecasting exercises.) The simple regression estimates are

$$\widehat{inf}_t = 1.01 + .505\, unem_t \quad [10.14]$$
$$(1.49)\ \ (.257)$$
$$n = 59,\ R^2 = .065,\ \overline{R}^2 = .049.$$

This equation does not suggest a tradeoff between *unem* and *inf*: $\hat{\beta}_1 > 0$. The t statistic for $\hat{\beta}_1$ is about 1.96, which gives a p-value against a two-sided alternative of about .055. Thus, if anything, there is a positive relationship between inflation and unemployment.

There are some problems with this analysis that we cannot address in detail now. In Chapter 12, we will see that the CLM assumptions do not hold. In addition, the static Phillips curve is probably not the best model for determining whether there is a short-run tradeoff between inflation and unemployment. Macroeconomists generally prefer the expectations augmented Phillips curve, a simple example of which is given in Chapter 11.

As a second example, we estimate equation (10.11) using annual data on the U.S. economy.

EXAMPLE 10.2 **Effects of Inflation and Deficits on Interest Rates**

The data in INTDEF come from the 2004 *Economic Report of the President* (Tables B-73 and B-79) and span the years 1948 through 2003. The variable *i3* is the three-month T-bill rate, *inf* is the annual inflation rate based on the consumer price index (CPI), and *def* is the federal budget deficit as a percentage of GDP. The estimated equation is

$$\widehat{i3}_t = 1.73 + .606\, inf_t + .513\, def_t \quad [10.15]$$
$$(0.43)\ (.082)\qquad (.118)$$
$$n = 56,\ R^2 = .602,\ \overline{R}^2 = .587.$$

These estimates show that increases in inflation or the relative size of the deficit increase short-term interest rates, both of which are expected from basic economics. For example, a ceteris paribus one percentage point increase in the inflation rate increases *i3* by .606 points. Both *inf* and *def* are very statistically significant, assuming, of course, that the CLM assumptions hold.

10-4 Functional Form, Dummy Variables, and Index Numbers

All of the functional forms we learned about in earlier chapters can be used in time series regressions. The most important of these is the natural logarithm: time series regressions with constant percentage effects appear often in applied work.

EXAMPLE 10.3 **Puerto Rican Employment and the Minimum Wage**

Annual data on the Puerto Rican employment rate, minimum wage, and other variables are used by Castillo-Freeman and Freeman (1992) to study the effects of the U.S. minimum wage on employment in Puerto Rico. A simplified version of their model is

$$\log(prepop_t) = \beta_0 + \beta_1 \log(mincov_t) + \beta_2 \log(usgnp_t) + u_t, \quad [10.16]$$

where $prepop_t$ is the employment rate in Puerto Rico during year t (ratio of those working to total population), $usgnp_t$ is real U.S. gross national product (in billions of dollars), and *mincov* measures the importance of the minimum wage relative to average wages. In particular, *mincov* = (*avgmin*/*avgwage*)·*avgcov*, where *avgmin* is the average minimum wage, *avgwage* is the average overall wage, and *avgcov* is the average coverage rate (the proportion of workers actually covered by the minimum wage law).

Using the data in PRMINWGE for the years 1950 through 1987 gives

$$\widehat{\log(prepop_t)} = -1.05 - .154\log(mincov_t) - .012\log(usgnp_t)$$
$$(0.77)\ (.065)\qquad (.089) \qquad [10.17]$$
$$n = 38, R^2 = .661, \overline{R}^2 = .641.$$

The estimated elasticity of *prepop* with respect to *mincov* is $-.154$, and it is statistically significant with $t = -2.37$. Therefore, a higher minimum wage lowers the employment rate, something that classical economics predicts. The GNP variable is not statistically significant, but this changes when we account for a time trend in the next section.

We can use logarithmic functional forms in distributed lag models, too. For example, for quarterly data, suppose that money demand (M_t) and gross domestic product (GDP_t) are related by

$$\log(M_t) = \alpha_0 + \delta_0\log(GDP_t) + \delta_1\log(GDP_{t-1}) + \delta_2\log(GDP_{t-2}) + \delta_3\log(GDP_{t-3}) + \delta_4\log(GDP_{t-4}) + u_t.$$

The impact propensity in this equation, δ_0, is also called the **short-run elasticity**: it measures the immediate percentage change in money demand given a 1% increase in *GDP*. The LRP, $\delta_0 + \delta_1 + \cdots + \delta_4$, is sometimes called the **long-run elasticity**: it measures the percentage increase in money demand after four quarters given a permanent 1% increase in *GDP*.

Binary or dummy independent variables are also quite useful in time series applications. Because the unit of observation is time, a dummy variable represents whether, in each time period, a certain event has occurred. For example, for annual data, we can indicate in each year whether a Democrat or a Republican is president of the United States by defining a variable $democ_t$, which is unity if the president is a Democrat, and zero otherwise. Or, in looking at the effects of capital punishment on murder rates in Texas, we can define a dummy variable for each year equal to one if Texas had capital punishment during that year, and zero otherwise.

Often, dummy variables are used to isolate certain periods that may be systematically different from other periods covered by a data set.

EXAMPLE 10.4 Effects of Personal Exemption on Fertility Rates

The general fertility rate (*gfr*) is the number of children born to every 1,000 women of childbearing age. For the years 1913 through 1984, the equation,

$$gfr_t = \beta_0 + \beta_1 pe_t + \beta_2 ww2_t + \beta_3 pill_t + u_t,$$

explains *gfr* in terms of the average real dollar value of the personal tax exemption (*pe*) and two binary variables. The variable *ww2* takes on the value unity during the years 1941 through 1945, when the United States was involved in World War II. The variable *pill* is unity from 1963 onward, when the birth control pill was made available for contraception.

Using the data in FERTIL3, which were taken from the article by Whittington, Alm, and Peters (1990)

$$\widehat{gfr_t} = 98.68 + .083\, pe_t - 24.24\, ww2_t - 31.59\, pill_t$$
$$(3.21)\ (.030)\qquad (7.46)\qquad (4.08) \qquad [10.18]$$
$$n = 72, R^2 = .473, \overline{R}^2 = .450.$$

Each variable is statistically significant at the 1% level against a two-sided alternative. We see that the fertility rate was lower during World War II: given *pe*, there were about 24 fewer births for every 1,000 women of childbearing age, which is a large reduction. (From 1913 through 1984, *gfr* ranged from about 65 to 127.) Similarly, the fertility rate has been substantially lower since the introduction of the birth control pill.

The variable of economic interest is *pe*. The average *pe* over this time period is $100.40, ranging from zero to $243.83. The coefficient on *pe* implies that a $12.00 increase in *pe* increases *gfr* by about one birth per 1,000 women of childbearing age. This effect is hardly trivial.

In Section 10-2, we noted that the fertility rate may react to changes in *pe* with a lag. Estimating a distributed lag model with two lags gives

$$\widehat{gfr}_t = 95.87 + .073\,pe_t - .0058\,pe_{t-1} + .034\,pe_{t-2} - 22.12\,ww2_t - 31.30\,pill_t \quad [10.19]$$
$$(3.28)\ (.126)\quad (.1557)\quad (.126)\quad (10.73)\quad (3.98)$$
$$n = 70,\ R^2 = .499,\ \overline{R}^2 = .459.$$

In this regression, we only have 70 observations because we lose two when we lag *pe* twice. The coefficients on the *pe* variables are estimated very imprecisely, and each one is individually insignificant. It turns out that there is substantial correlation between pe_t, pe_{t-1}, and pe_{t-2}, and this multicollinearity makes it difficult to estimate the effect at each lag. However, pe_t, pe_{t-1}, and pe_{t-2} are jointly significant: the *F* statistic has a *p*-value = .012. Thus, *pe* does have an effect on *gfr* [as we already saw in (10.18)], but we do not have good enough estimates to determine whether it is contemporaneous or with a one- or two-year lag (or some of each). Actually, pe_{t-1} and pe_{t-2} are jointly insignificant in this equation (p-value = .95), so at this point, we would be justified in using the static model. But for illustrative purposes, let us obtain a confidence interval for the LRP in this model.

The estimated LRP in (10.19) is .073 − .0058 + .034 ≈ .101. However, we do not have enough information in (10.19) to obtain the standard error of this estimate. To obtain the standard error of the estimated LRP, we use the trick suggested in Section 4-4. Let $\theta_0 = \delta_0 + \delta_1 + \delta_2$ denote the LRP and write δ_0 in terms of θ_0, δ_1, and δ_2 as $\delta_0 = \theta_0 - \delta_1 - \delta_2$. Next, substitute for δ_0 in the model

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \cdots$$

to get

$$\begin{aligned} gfr_t &= \alpha_0 + (\theta_0 - \delta_1 - \delta_2)pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \cdots \\ &= \alpha_0 + \theta_0 pe_t + \delta_1(pe_{t-1} - pe_t) + \delta_2(pe_{t-2} - pe_t) + \cdots. \end{aligned}$$

From this last equation, we can obtain $\hat{\theta}_0$ and its standard error by regressing gfr_t on pe_t, $(pe_{t-1} - pe_t)$, $(pe_{t-2} - pe_t)$, $ww2_t$, and $pill_t$. The coefficient and associated standard error on pe_t are what we need. Running this regression gives $\hat{\theta}_0 = .101$ as the coefficient on pe_t (as we already knew) and $\text{se}(\hat{\theta}_0) = .030$ [which we could not compute from (10.19)]. Therefore, the *t* statistic for $\hat{\theta}_0$ is about 3.37, so $\hat{\theta}_0$ is statistically different from zero at small significance levels. Even though none of the $\hat{\delta}_j$ is individually significant, the LRP is very significant. The 95% confidence interval for the LRP is about .041 to .160.

Whittington, Alm, and Peters (1990) allow for further lags but restrict the coefficients to help alleviate the multicollinearity problem that hinders estimation of the individual δ_j. (See Problem 6 for an example of how to do this.) For estimating the LRP, which would seem to be of primary interest here, such restrictions are unnecessary. Whittington, Alm, and Peters also control for additional variables, such as average female wage and the unemployment rate.

Binary explanatory variables are the key component in what is called an **event study**. In an event study, the goal is to see whether a particular event influences some outcome. Economists who study industrial organization have looked at the effects of certain events on firm stock prices. For example, Rose (1985) studied the effects of new trucking regulations on the stock prices of trucking companies.

A simple version of an equation used for event studies is

$$R_t^f = \beta_0 + \beta_1 R_t^m + \beta_2 d_t + u_t,$$

where R_t^f is the stock return for firm f during period t (usually a week or a month), R_t^m is the market return (usually computed for a broad stock market index), and d_t is a dummy variable indicating when the event occurred. For example, if the firm is an airline, d_t might denote whether the airline experienced a publicized accident or near accident during week t. Including R_t^m in the equation controls for the possibility that broad market movements might coincide with airline accidents. Sometimes, multiple dummy variables are used. For example, if the event is the imposition of a new regulation that might affect a certain firm, we might include a dummy variable that is one for a few weeks before the regulation was publicly announced and a second dummy variable for a few weeks after the regulation was announced. The first dummy variable might detect the presence of inside information.

Before we give an example of an event study, we need to discuss the notion of an **index number** and the difference between nominal and real economic variables. An index number typically aggregates a vast amount of information into a single quantity. Index numbers are used regularly in time series analysis, especially in macroeconomic applications. An example of an index number is the index of industrial production (IIP), computed monthly by the Board of Governors of the Federal Reserve. The IIP is a measure of production across a broad range of industries, and, as such, its magnitude in a particular year has no quantitative meaning. In order to interpret the magnitude of the IIP, we must know the **base period** and the **base value**. In the 1997 *Economic Report of the President* (*ERP*), the base year is 1987, and the base value is 100. (Setting IIP to 100 in the base period is just a convention; it makes just as much sense to set IIP = 1 in 1987, and some indexes are defined with 1 as the base value.) Because the IIP was 107.7 in 1992, we can say that industrial production was 7.7% higher in 1992 than in 1987. We can use the IIP in any two years to compute the percentage difference in industrial output during those two years. For example, because IIP = 61.4 in 1970 and IIP = 85.7 in 1979, industrial production grew by about 39.6% during the 1970s.

It is easy to change the base period for any index number, and sometimes we must do this to give index numbers reported with different base years a common base year. For example, if we want to change the base year of the IIP from 1987 to 1982, we simply divide the IIP for each year by the 1982 value and then multiply by 100 to make the base period value 100. Generally, the formula is

$$newindex_t = 100(oldindex_t/oldindex_{\text{newbase}}), \qquad [10.20]$$

where $oldindex_{\text{newbase}}$ is the original value of the index in the new base year. For example, with base year 1987, the IIP in 1992 is 107.7; if we change the base year to 1982, the IIP in 1992 becomes $100(107.7/81.9) = 131.5$ (because the IIP in 1982 was 81.9).

Another important example of an index number is a *price index*, such as the CPI. We already used the CPI to compute annual inflation rates in Example 10.1. As with the industrial production index, the CPI is only meaningful when we compare it across different years (or months, if we are using monthly data). In the 1997 *ERP*, CPI = 38.8 in 1970 and CPI = 130.7 in 1990. Thus, the general price level grew by almost 237% over this 20-year period. (In 1997, the CPI is defined so that its average in 1982, 1983, and 1984 equals 100; thus, the base period is listed as 1982–1984.)

In addition to being used to compute inflation rates, price indexes are necessary for turning a time series measured in *nominal dollars* (or *current dollars*) into *real dollars* (or *constant dollars*). Most economic behavior is assumed to be influenced by real, not nominal, variables. For example, classical labor economics assumes that labor supply is based on the real hourly wage, not the nominal wage. Obtaining the real wage from the nominal wage is easy if we have a price index such as the CPI. We must be a little careful to first divide the CPI by 100, so that the value in the base year is 1. Then, if w denotes the average hourly wage in nominal dollars and p = CPI/100, the *real wage* is simply w/p. This wage is measured in dollars for the base period of the CPI. For example, in Table B-45 in the 1997 *ERP*, average hourly earnings are reported in nominal terms and in 1982 dollars (which means that the CPI used in computing the real wage had the base year 1982). This table reports that the nominal hourly wage in 1960 was \$2.09, but measured in 1982 dollars, the wage was \$6.79. The real hourly wage had peaked in 1973, at \$8.55 in 1982 dollars, and had fallen to \$7.40 by 1995. Thus,

there was a nontrivial decline in real wages over those 22 years. (If we compare nominal wages from 1973 and 1995, we get a very misleading picture: \$3.94 in 1973 and \$11.44 in 1995. Because the real wage fell, the increase in the nominal wage was due entirely to inflation.)

Standard measures of economic output are in real terms. The most important of these is *gross domestic product*, or *GDP*. When growth in GDP is reported in the popular press, it is always *real* GDP growth. In the 2012 *ERP*, Table B-2, GDP is reported in billions of 2005 dollars. We used a similar measure of output, real gross national product, in Example 10.3.

Interesting things happen when real dollar variables are used in combination with natural logarithms. Suppose, for example, that average weekly hours worked are related to the real wage as

$$\log(hours) = \beta_0 + \beta_1 \log(w/p) + u.$$

Using the fact that $\log(w/p) = \log(w) - \log(p)$, we can write this as

$$\log(hours) = \beta_0 + \beta_1 \log(w) + \beta_2 \log(p) + u, \qquad [10.21]$$

but with the restriction that $\beta_2 = -\beta_1$. Therefore, the assumption that only the real wage influences labor supply imposes a restriction on the parameters of model (10.21). If $\beta_2 \neq -\beta_1$, then the price level has an effect on labor supply, something that can happen if workers do not fully understand the distinction between real and nominal wages.

There are many practical aspects to the actual computation of index numbers, but it would take us too far afield to cover those here. Detailed discussions of price indexes can be found in most intermediate macroeconomic texts, such as Mankiw (1994, Chapter 2). For us, it is important to be able to use index numbers in regression analysis. As mentioned earlier, because the magnitudes of index numbers are not especially informative, they often appear in logarithmic form, so that regression coefficients have percentage change interpretations.

We now give an example of an event study that also uses index numbers.

EXAMPLE 10.5 Antidumping Filings and Chemical Imports

Krupp and Pollard (1996) analyzed the effects of antidumping filings by U.S. chemical industries on imports of various chemicals. We focus here on one industrial chemical, barium chloride, a cleaning agent used in various chemical processes and in gasoline production. The data are contained in the file BARIUM. In the early 1980s, U.S. barium chloride producers believed that China was offering its U.S. imports an unfairly low price (an action known as *dumping*), and the barium chloride industry filed a complaint with the U.S. International Trade Commission (ITC) in October 1983. The ITC ruled in favor of the U.S. barium chloride industry in October 1984. There are several questions of interest in this case, but we will touch on only a few of them. First, were imports unusually high in the period immediately preceding the initial filing? Second, did imports change noticeably after an antidumping filing? Finally, what was the reduction in imports after a decision in favor of the U.S. industry?

To answer these questions, we follow Krupp and Pollard by defining three dummy variables: *befile6* is equal to 1 during the six months before filing, *affile6* indicates the six months after filing, and *afdec6* denotes the six months after the positive decision. The dependent variable is the volume of imports of barium chloride from China, *chnimp*, which we use in logarithmic form. We include as explanatory variables, all in logarithmic form, an index of chemical production, *chempi* (to control for overall demand for barium chloride), the volume of gasoline production, *gas* (another demand variable), and an exchange rate index, *rtwex*, which measures the strength of the dollar against several other currencies. The chemical production index was defined to be 100 in June 1977. The analysis here differs somewhat from Krupp and Pollard in that we use natural logarithms of all variables (except the dummy variables, of course), and we include all three dummy variables in the same regression.

Using monthly data from February 1978 through December 1988 gives the following:

$$\widehat{\log(chnimp)} = -17.80 + 3.12\log(chempi) + .196\log(gas) \quad (21.05) \quad (.48) \quad (.907)$$
$$+ .983\log(rtwex) + .060\,befile6 - .032\,affile6 - .565\,afdec6 \quad [10.22] \quad (.400) \quad (.261) \quad (.264) \quad (.286)$$
$$n = 131, R^2 = .305, \overline{R}^2 = .271.$$

The equation shows that *befile6* is statistically insignificant, so there is no evidence that Chinese imports were unusually high during the six months before the suit was filed. Further, although the estimate on *affile6* is negative, the coefficient is small (indicating about a 3.2% fall in Chinese imports), and it is statistically very insignificant. The coefficient on *afdec6* shows a substantial fall in Chinese imports of barium chloride after the decision in favor of the U.S. industry, which is not surprising. Because the effect is so large, we compute the exact percentage change: $100[\exp(-.565) - 1] \approx -43.2\%$. The coefficient is statistically significant at the 5% level against a two-sided alternative.

The coefficient signs on the control variables are what we expect: an increase in overall chemical production increases the demand for the cleaning agent. Gasoline production does not affect Chinese imports significantly. The coefficient on log(*rtwex*) shows that an increase in the value of the dollar relative to other currencies increases the demand for Chinese imports, as is predicted by economic theory. (In fact, the elasticity is not statistically different from 1. Why?)

Interactions among qualitative and quantitative variables are also used in time series analysis. An example with practical importance follows.

EXAMPLE 10.6 Election Outcomes and Economic Performance

Fair (1996) summarizes his work on explaining presidential election outcomes in terms of economic performance. He explains the proportion of the two-party vote going to the Democratic candidate using data for the years 1916 through 1992 (every four years) for a total of 20 observations. We estimate a simplified version of Fair's model (using variable names that are more descriptive than his):

$$demvote = \beta_0 + \beta_1 partyWH + \beta_2 incum + \beta_3 partyWH\cdot gnews + \beta_4 partyWH\cdot inf + u,$$

where *demvote* is the proportion of the two-party vote going to the Democratic candidate. The explanatory variable *partyWH* is similar to a dummy variable, but it takes on the value 1 if a Democrat is in the White House and −1 if a Republican is in the White House. Fair uses this variable to impose the restriction that the effects of a Republican or a Democrat being in the White House have the same magnitude but the opposite sign. This is a natural restriction because the party shares must sum to one, by definition. It also saves two degrees of freedom, which is important with so few observations. Similarly, the variable *incum* is defined to be 1 if a Democratic incumbent is running, −1 if a Republican incumbent is running, and zero otherwise. The variable *gnews* is the number of quarters, during the administration's first 15 quarters, when the quarterly growth in real per capita output was above 2.9% (at an annual rate), and *inf* is the average annual inflation rate over the first 15 quarters of the administration. See Fair (1996) for precise definitions.

Economists are most interested in the interaction terms *partyWH·gnews* and *partyWH·inf*. Because *partyWH* equals 1 when a Democrat is in the White House, β_3 measures the effect of good economic news on the party in power; we expect $\beta_3 > 0$. Similarly, β_4 measures the effect that inflation has on the party in power. Because inflation during an administration is considered to be bad news, we expect $\beta_4 < 0$.

The estimated equation using the data in FAIR is

$$\widehat{demvote} = .481 - .0435\,partyWH + .0544\,incum \\ (.012)\ (.0405)\qquad (.0234) \\ + .0108\,partyWH \cdot gnews - .0077\,partyWH \cdot inf \qquad [10.23] \\ (.0041)\qquad (.0033) \\ n = 20, R^2 = .663, \bar{R}^2 = .573.$$

All coefficients, except that on *partyWH*, are statistically significant at the 5% level. Incumbency is worth about 5.4 percentage points in the share of the vote. (Remember, *demvote* is measured as a proportion.) Further, the economic news variable has a positive effect: one more quarter of good news is worth about 1.1 percentage points. Inflation, as expected, has a negative effect: if average annual inflation is, say, two percentage points higher, the party in power loses about 1.5 percentage points of the two-party vote.

We could have used this equation to predict the outcome of the 1996 presidential election between Bill Clinton, the Democrat, and Bob Dole, the Republican. (The independent candidate, Ross Perot, is excluded because Fair's equation is for the two-party vote only.) Because Clinton ran as an incumbent, *partyWH* = 1 and *incum* = 1. To predict the election outcome, we need the variables *gnews* and *inf*. During Clinton's first 15 quarters in office, the annual growth rate of per capita real GDP exceeded 2.9% three times, so *gnews* = 3. Further, using the GDP price deflator reported in Table B-4 in the 1997 *ERP*, the average annual inflation rate (computed using Fair's formula) from the fourth quarter in 1991 to the third quarter in 1996 was 3.019. Plugging these into (10.23) gives

$$\widehat{demvote} = .481 - .0435 + .0544 + .0108(3) - .0077(3.019) \approx .5011.$$

Therefore, based on information known before the election in November, Clinton was predicted to receive a very slight majority of the two-party vote: about 50.1%. In fact, Clinton won more handily: his share of the two-party vote was 54.65%.

10-5 Trends and Seasonality

10-5a Characterizing Trending Time Series

Many economic time series have a common tendency of growing over time. We must recognize that some series contain a **time trend** in order to draw causal inference using time series data. Ignoring the fact that two sequences are trending in the same or opposite directions can lead us to falsely conclude that changes in one variable are actually caused by changes in another variable. In many cases, two time series processes appear to be correlated only because they are both trending over time for reasons related to other unobserved factors.

Figure 10.2 contains a plot of labor productivity (output per hour of work) in the United States for the years 1947 through 1987. This series displays a clear upward trend, which reflects the fact that workers have become more productive over time.

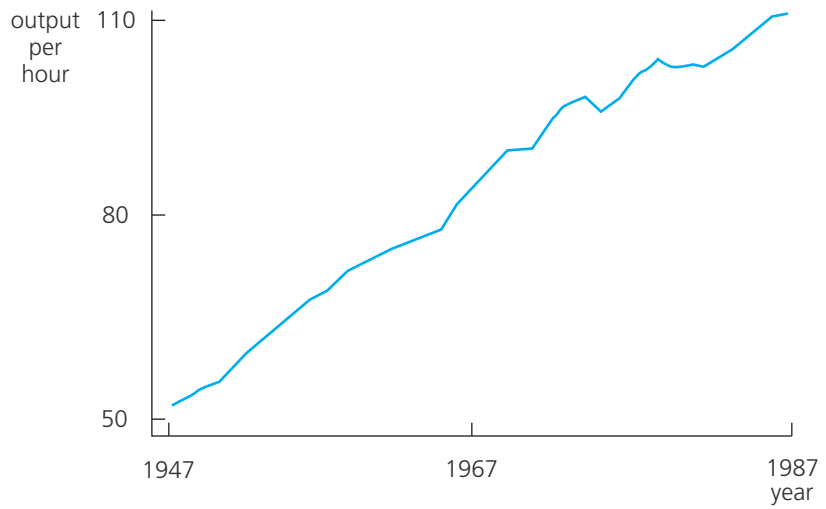
Other series, at least over certain time periods, have clear downward trends. Because positive trends are more common, we will focus on those during our discussion.

What kind of statistical models adequately capture trending behavior? One popular formulation is to write the series $\{y_t\}$ as

$$y_t = \alpha_0 + \alpha_1 t + e_t, t = 1, 2, \ldots, \qquad [10.24]$$

where, in the simplest case, $\{e_t\}$ is an independent, identically distributed (i.i.d.) sequence with $\mathrm{E}(e_t) = 0$ and $\mathrm{Var}(e_t) = \sigma_e^2$. Note how the parameter α_1 multiplies time, t, resulting in a **linear time trend**.

FIGURE 10.2 Output per labor hour in the United States during the years 1947–1987; 1977 = 100.



Interpreting α_1 in (10.24) is simple: holding all other factors (those in e_1) fixed, α_1 measures the change in y_t from one period to the next due to the passage of time. We can write this mathematically by defining the change in e_t from period $t-1$ to t as $\Delta e_t = e_t - e_{t-1}$. Equation (10.24) implies that if $\Delta e_t = 0$ then

$$\Delta y_t = y_t - y_{t-1} = \alpha_1.$$

Another way to think about a sequence that has a linear time trend is that its average value is a linear function of time:

$$\mathrm{E}(y_t) = \alpha_0 + \alpha_1 t. \qquad [10.25]$$

If $\alpha_1 > 0$, then, on average, y_t is growing over time and therefore has an upward trend. If $\alpha_1 < 0$, then y_t has a downward trend. The values of y_t do not fall exactly on the line in (10.25) due to randomness, but the expected values are on the line. Unlike the mean, the variance of y_t is constant across time: $\mathrm{Var}(y_t) = \mathrm{Var}(e_t) = \sigma_e^2$.

GOING FURTHER 10.4

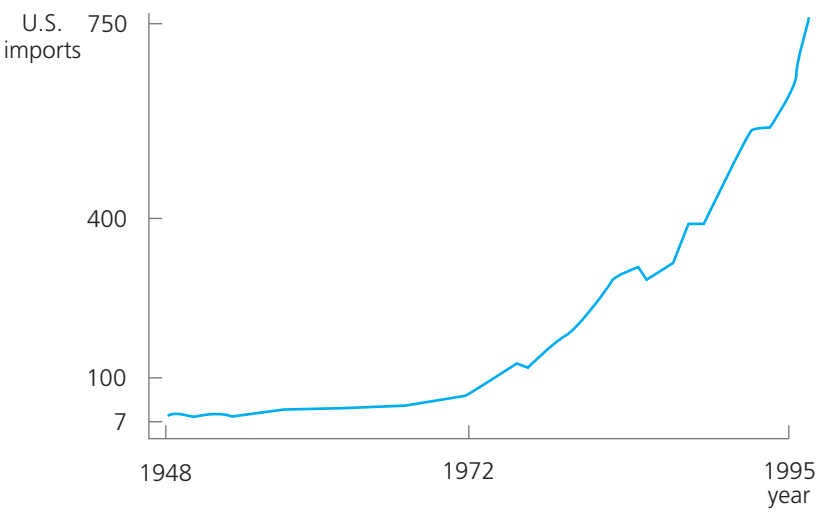
In Example 10.4, we used the general fertility rate as the dependent variable in an FDL model. From 1950 through the mid-1980s, the *gfr* has a clear downward trend. Can a linear trend with $\alpha_1 < 0$ be realistic for all future time periods? Explain.

If $\{e_t\}$ is an i.i.d. sequence, then $\{y_t\}$ is an independent, though not identically, distributed sequence. A more realistic characterization of trending time series allows $\{e_t\}$ to be correlated over time, but this does not change the flavor of a linear time trend. In fact, what is important for regression analysis under the classical linear model assumptions is that $\mathrm{E}\{y_t\}$ is linear in t. When we cover large sample properties of OLS in Chapter 11, we will have to discuss how much temporal correlation in $\{e_t\}$ is allowed.

Many economic time series are better approximated by an **exponential trend**, which follows when a series has the same average growth rate from period to period. Figure 10.3 plots data on annual nominal imports for the United States during the years 1948 through 1995 (*ERP* 1997, Table B-101).

In the early years, we see that the change in imports over each year is relatively small, whereas the change increases as time passes. This is consistent with a *constant average growth rate*: the percentage change is roughly the same in each period.

FIGURE 10.3 Nominal U.S. imports during the years 1948–1995 (in billions of U.S. dollars).



In practice, an exponential trend in a time series is captured by modeling the natural logarithm of the series as a linear trend (assuming that $y_t > 0$):

$$\log(y_t) = \beta_0 + \beta_1 t + e_t, \; t = 1, 2, \ldots . \qquad [10.26]$$

Exponentiating shows that y_t itself has an exponential trend: $y_t = \exp(\beta_0 + \beta_1 t + e_t)$. Because we will want to use exponentially trending time series in linear regression models, (10.26) turns out to be the most convenient way for representing such series.

How do we interpret β_1 in (10.26)? Remember that, for small changes, $\Delta\log(y_t) = \log(y_t) - \log(y_{t-1})$ is approximately the proportionate change in y_t:

$$\Delta\log(y_t) \approx (y_t - y_{t-1})/y_{t-1}. \qquad [10.27]$$

The right-hand side of (10.27) is also called the **growth rate** in y from period $t-1$ to period t. To turn the growth rate into a percentage, we simply multiply by 100. If y_t follows (10.26), then, taking changes and setting $\Delta e_t = 0$,

$$\Delta\log(y_t) = \beta_1, \text{ for all } t. \qquad [10.28]$$

In other words, β_1 is approximately the average per period growth rate in y_t. For example, if t denotes year and $\beta_1 = .027$, then y_t grows about 2.7% per year on average.

Although linear and exponential trends are the most common, time trends can be more complicated. For example, instead of the linear trend model in (10.24), we might have a quadratic time trend:

$$y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t. \qquad [10.29]$$

If α_1 and α_2 are positive, then the slope of the trend is increasing, as is easily seen by computing the approximate slope (holding e_t fixed):

$$\frac{\Delta y_t}{\Delta t} \approx \alpha_1 + 2\alpha_2 t. \qquad [10.30]$$

[If you are familiar with calculus, you recognize the right-hand side of (10.30) as the derivative of $\alpha_0 + \alpha_1 t + \alpha_2 t^2$ with respect to t.] If $\alpha_1 > 0$, but $\alpha_2 < 0$, the trend has a hump shape. This may not be a very good description of certain trending series because it requires an increasing trend to be followed, eventually, by a decreasing trend. Nevertheless, over a given time span, it can be a flexible way of modeling time series that have more complicated trends than either (10.24) or (10.26).

10-5b Using Trending Variables in Regression Analysis

Accounting for explained or explanatory variables that are trending is fairly straightforward in regression analysis. First, nothing about trending variables necessarily violates the classical linear model Assumptions TS.1 through TS.6. However, we must be careful to allow for the fact that unobserved, trending factors that affect y_t might also be correlated with the explanatory variables. If we ignore this possibility, we may find a spurious relationship between y_t and one or more explanatory variables. The phenomenon of finding a relationship between two or more trending variables simply because each is growing over time is an example of a **spurious regression problem**. Fortunately, adding a time trend eliminates this problem.

For concreteness, consider a model where two observed factors, x_{t1} and x_{t2}, affect y_t. In addition, there are unobserved factors that are systematically growing or shrinking over time. A model that captures this is

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 t + u_t. \qquad [10.31]$$

This fits into the multiple linear regression framework with $x_{t3} = t$. Allowing for the trend in this equation explicitly recognizes that y_t may be growing $(\beta_3 > 0)$ or shrinking $(\beta_3 < 0)$ over time for reasons essentially unrelated to x_{t1} and x_{t2}. If (10.31) satisfies assumptions TS.1, TS.2, and TS.3, then omitting t from the regression and regressing y_t on x_{t1}, x_{t2} will generally yield biased estimators of β_1 and β_2: we have effectively omitted an important variable, t, from the regression. This is especially true if x_{t1} and x_{t2} are themselves trending, because they can then be highly correlated with t. The next example shows how omitting a time trend can result in spurious regression.

EXAMPLE 10.7 Housing Investment and Prices

The data in HSEINV are annual observations on housing investment and a housing price index in the United States for 1947 through 1988. Let *invpc* denote real per capita housing investment (in thousands of dollars) and let *price* denote a housing price index (equal to 1 in 1982). A simple regression in constant elasticity form, which can be thought of as a supply equation for housing stock, gives

$$\widehat{\log(invpc)} = -.550 + 1.241 \log(price) \qquad [10.32]$$
$$(.043) \quad (.382)$$
$$n = 42, R^2 = .208, \bar{R}^2 = .189.$$

The elasticity of per capita investment with respect to price is very large and statistically significant; it is not statistically different from one. We must be careful here. Both *invpc* and *price* have upward trends. In particular, if we regress $\log(invpc)$ on t, we obtain a coefficient on the trend equal to .0081 (standard error = .0018); the regression of $\log(price)$ on t yields a trend coefficient equal to .0044 (standard error = .0004). Although the standard errors on the trend coefficients are not necessarily reliable—these regressions tend to contain substantial serial correlation—the coefficient estimates do reveal upward trends.

To account for the trending behavior of the variables, we add a time trend:

$$\widehat{\log(invpc)} = -.913 - .381\log(price) + .0098\,t \quad [10.33]$$
$$(1.36)\quad (.679)\qquad\qquad (.0035)$$
$$n = 42,\ R^2 = .341,\ \bar{R}^2 = .307.$$

The story is much different now: the estimated price elasticity is negative and not statistically different from zero. The time trend is statistically significant, and its coefficient implies an approximate 1% increase in *invpc* per year, on average. From this analysis, we cannot conclude that real per capita housing investment is influenced at all by price. There are other factors, captured in the time trend, that affect *invpc*, but we have not modeled these. The results in (10.32) show a spurious relationship between *invpc* and *price* due to the fact that price is also trending upward over time.

In some cases, adding a time trend can make a key explanatory variable *more* significant. This can happen if the dependent and independent variables have different kinds of trends (say, one upward and one downward), but movement in the independent variable *about* its trend line causes movement in the dependent variable away from its trend line.

EXAMPLE 10.8 **Fertility Equation**

If we add a linear time trend to the fertility equation (10.18), we obtain

$$\widehat{gfr}_t = 111.77 + .279\,pe_t - 35.59\,ww2_t + .997\,pill_t - 1.15\,t \quad [10.34]$$
$$(3.36)\ (.040)\qquad (6.30)\qquad (6.626)\qquad (.19)$$
$$n = 72,\ R^2 = .662,\ \bar{R}^2 = .642.$$

The coefficient on *pe* is more than triple the estimate from (10.18), and it is much more statistically significant. Interestingly, *pill* is not significant once an allowance is made for a linear trend. As can be seen by the estimate, *gfr* was falling, on average, over this period, other factors being equal.

Because the general fertility rate exhibited both upward and downward trends during the period from 1913 through 1984, we can see how robust the estimated effect of *pe* is when we use a quadratic trend:

$$\widehat{gfr}_t = 124.09 + .348\,pe_t - 35.88\,ww2_t - 10.12\,pill_t - 2.53\,t + .0196\,t^2 \quad [10.35]$$
$$(4.36)\ (.040)\qquad (5.71)\qquad (6.34)\qquad (.39)\ \ (.0050)$$
$$n = 72,\ R^2 = .727,\ \bar{R}^2 = .706.$$

The coefficient on *pe* is even larger and more statistically significant. Now, *pill* has the expected negative effect and is marginally significant, and both trend terms are statistically significant. The quadratic trend is a flexible way to account for the unusual trending behavior of *gfr*.

You might be wondering in Example 10.8: why stop at a quadratic trend? Nothing prevents us from adding, say, t^3 as an independent variable, and, in fact, this might be warranted (see Computer Exercise C6). But we have to be careful not to get carried away when including trend terms in a model. We want relatively simple trends that capture broad movements in the dependent variable that are not explained by the independent variables in the model. If we include enough polynomial terms in t, then we can track any series pretty well. But this offers little help in finding which explanatory variables affect y_t.

10-5c A Detrending Interpretation of Regressions with a Time Trend

Including a time trend in a regression model creates a nice interpretation in terms of **detrending** the original data series before using them in regression analysis. For concreteness, we focus on model (10.31), but our conclusions are much more general.

When we regress y_t on x_{t1}, x_{t2}, and t, we obtain the fitted equation

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \hat{\beta}_2 x_{t2} + \hat{\beta}_3 t. \qquad [10.36]$$

We can extend the Frisch-Waugh result on the partialling out interpretation of OLS that we covered in Section 3-2 to show that $\hat{\beta}_1$ and $\hat{\beta}_2$ can be obtained as follows.

(i) Regress each of y_t, x_{t1}, and x_{t2} on a constant and the time trend t and save the residuals, say, $\ddot{y}_t$, $\ddot{x}_{t1}$, $\ddot{x}_{t2}$, $t = 1, 2, \ldots, n$. For example,

$$\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 t.$$

Thus, we can think of $\ddot{y}_t$ as being *linearly detrended.* In detrending y_t, we have estimated the model

$$y_t = \alpha_0 + \alpha_1 t + e_t$$

by OLS; the residuals from this regression, $\hat{e}_t = \ddot{y}_t$, have the time trend removed (at least in the sample). A similar interpretation holds for $\ddot{x}_{t1}$ and $\ddot{x}_{t2}$.

(ii) Run the regression of

$$\ddot{y}_t \text{ on } \ddot{x}_{t1}, \ddot{x}_{t2}. \qquad [10.37]$$

(No intercept is necessary, but including an intercept affects nothing: the intercept will be estimated to be zero.) This regression exactly yields $\hat{\beta}_1$ and $\hat{\beta}_2$ from (10.36).

This means that the estimates of primary interest, $\hat{\beta}_1$ and $\hat{\beta}_2$, can be interpreted as coming from a regression *without* a time trend, but where we first detrend the dependent variable and all other independent variables. The same conclusion holds with any number of independent variables and if the trend is quadratic or of some other polynomial degree.

If t is omitted from (10.36), then no detrending occurs, and y_t might seem to be related to one or more of the x_{tj} simply because each contains a trend; we saw this in Example 10.7. If the trend term is statistically significant, and the results change in important ways when a time trend is added to a regression, then the initial results without a trend should be treated with suspicion.

The interpretation of $\hat{\beta}_1$ and $\hat{\beta}_2$ shows that it is a good idea to include a trend in the regression if any independent variable is trending, even if y_t is not. If y_t has no noticeable trend, but, say, x_{t1} is growing over time, then excluding a trend from the regression may make it look as if x_{t1} has no effect on y_t, even though movements of x_{t1} about its trend may affect y_t. This will be captured if t is included in the regression.

EXAMPLE 10.9 Puerto Rican Employment

When we add a linear trend to equation (10.17), the estimates are

$$\begin{aligned} \widehat{\log(prepop_t)} &= -8.70 - .169 \log(mincov_t) + 1.06 \log(usgnp_t) \\ &\qquad (1.30)\ \ (.044) \qquad\qquad\quad (0.18) \\ &\quad - .032\, t \\ &\qquad (.005) \end{aligned} \qquad [10.38]$$

$$n = 38,\ R^2 = .847,\ \bar{R}^2 = .834.$$

The coefficient on log(*usgnp*) has changed dramatically, from −.012 and insignificant to 1.06 and very significant. The coefficient on the minimum wage has changed only slightly, although the standard error is notably smaller, making log(*mincov*) more significant than before.

The variable $prepop_t$ displays no clear upward or downward trend, but log(*usgnp*) has an upward, linear trend. [A regression of log(*usgnp*) on t gives an estimate of about .03, so that *usgnp* is growing by about 3% per year over the period.] We can think of the estimate 1.06 as follows: when *usgnp* increases by 1% *above* its long-run trend, *prepop* increases by about 1.06%.

10-5d Computing *R*-Squared When the Dependent Variable Is Trending

R-squareds in time series regressions are often very high, especially compared with typical *R*-squareds for cross-sectional data. Does this mean that we learn more about factors affecting y from time series data? Not necessarily. On one hand, time series data often come in aggregate form (such as average hourly wages in the U.S. economy), and aggregates are often easier to explain than outcomes on individuals, families, or firms, which is often the nature of cross-sectional data. But the usual and adjusted *R*-squareds for time series regressions can be artificially high when the dependent variable is trending. Remember that R^2 is a measure of how large the error variance is relative to the variance of y. The formula for the adjusted *R*-squared shows this directly:

$$\bar{R}^2 = 1 - (\hat{\sigma}_u^2/\hat{\sigma}_y^2),$$

where $\hat{\sigma}_u^2$ is the unbiased estimator of the error variance, $\hat{\sigma}_y^2 = \text{SST}/(n-1)$, and $\text{SST} = \sum_{t=1}^{n}(y_t - \bar{y})^2$. Now, estimating the error variance when y_t is trending is no problem, provided a time trend is included in the regression. However, when $\text{E}(y_t)$ follows, say, a linear time trend [see (10.24)], $\text{SST}/(n-1)$ is no longer an unbiased or consistent estimator of $\text{Var}(y_t)$. In fact, $\text{SST}/(n-1)$ can substantially overestimate the variance in y_t, because it does not account for the trend in y_t.

When the dependent variable satisfies linear, quadratic, or any other polynomial trends, it is easy to compute a goodness-of-fit measure that first nets out the effect of any time trend on y_t. The simplest method is to compute the usual *R*-squared in a regression where the dependent variable has already been detrended. For example, if the model is (10.31), then we first regress y_t on t and obtain the residuals $\ddot{y}_t$. Then, we regress

$$\ddot{y}_t \text{ on } x_{t1}, x_{t2}, \text{ and } t. \qquad [10.39]$$

The *R*-squared from this regression is

$$1 - \frac{\text{SSR}}{\sum_{t=1}^{n}\ddot{y}_t^2}, \qquad [10.40]$$

where SSR is identical to the sum of squared residuals from (10.36). Because $\sum_{t=1}^{n}\ddot{y}_t^2 \leq \sum_{t=1}^{n}(y_t - \bar{y})^2$ (and usually the inequality is strict), the *R*-squared from (10.40) is no greater than, and usually less than, the *R*-squared from (10.36). (The sum of squared residuals is identical in both regressions.) When y_t contains a strong linear time trend, (10.40) can be much less than the usual *R*-squared.

The *R*-squared in (10.40) better reflects how well x_{t1} and x_{t2} explain y_t because it nets out the effect of the time trend. After all, we can always explain a trending variable with some sort of trend, but this does not mean we have uncovered any factors that cause movements in y_t. An adjusted *R*-squared can also be computed based on (10.40): divide SSR by $(n-4)$ because this is the *df* in (10.36) and divide $\sum_{t=1}^{n}\ddot{y}_t^2$ by $(n-2)$, as there are two trend parameters estimated in detrending y_t.

In general, SSR is divided by the *df* in the usual regression (that includes any time trends), and $\sum_{t=1}^{n} \ddot{y}_t^2$ is divided by $(n - p)$, where p is the number of trend parameters estimated in detrending y_t. Wooldridge (1991a) provides detailed suggestions for degrees-of-freedom corrections, but a computationally simple approach is fine as an approximation: use the adjusted *R*-squared from the regression $\ddot{y}_t$ on $t, t^2, \ldots, t^p, x_{t1}, \ldots, x_{tk}$. This requires us only to remove the trend from y_t to obtain $\ddot{y}_t$, and then we can use $\ddot{y}_t$ to compute the usual kinds of goodness-of-fit measures.

EXAMPLE 10.10 Housing Investment

In Example 10.7, we saw that including a linear time trend along with log(*price*) in the housing investment equation had a substantial effect on the price elasticity. But the *R*-squared from regression (10.33), taken literally, says that we are "explaining" 34.1% of the variation in log(*invpc*). This is misleading. If we first detrend log(*invpc*) and regress the detrended variable on log(*price*) and *t*, the *R*-squared becomes .008, and the adjusted *R*-squared is actually negative. Thus, movements in log(*price*) about its trend have virtually no explanatory power for movements in log(*invpc*) about its trend. This is consistent with the fact that the *t* statistic on log(*price*) in equation (10.33) is very small.

Before leaving this subsection, we must make a final point. In computing the *R*-squared form of an *F* statistic for testing multiple hypotheses, we just use the usual *R*-squareds without any detrending. Remember, the *R*-squared form of the *F* statistic is just a computational device, and so the usual formula is always appropriate.

10-5e Seasonality

If a time series is observed at monthly or quarterly intervals (or even weekly or daily), it may exhibit **seasonality**. For example, monthly housing starts in the Midwest are strongly influenced by weather. Although weather patterns are somewhat random, we can be sure that the weather during January will usually be more inclement than in June, and so housing starts are generally higher in June than in January. One way to model this phenomenon is to allow the expected value of the series, y_t, to be different in each month. As another example, retail sales in the fourth quarter are typically higher than in the previous three quarters because of the Christmas holiday. Again, this can be captured by allowing the average retail sales to differ over the course of a year. This is in addition to possibly allowing for a trending mean. For example, retail sales in the most recent first quarter were higher than retail sales in the fourth quarter from 30 years ago, because retail sales have been steadily growing. Nevertheless, if we compare average sales within a typical year, the seasonal holiday factor tends to make sales larger in the fourth quarter.

Even though many monthly and quarterly data series display seasonal patterns, not all of them do. For example, there is no noticeable seasonal pattern in monthly interest or inflation rates. In addition, series that do display seasonal patterns are often **seasonally adjusted** before they are reported for public use. A seasonally adjusted series is one that, in principle, has had the seasonal factors removed from it. Seasonal adjustment can be done in a variety of ways, and a careful discussion is beyond the scope of this text. [See Harvey (1990) and Hylleberg (1992) for detailed treatments.]

Seasonal adjustment has become so common that it is not possible to get seasonally unadjusted data in many cases. Quarterly U.S. GDP is a leading example. In the annual *Economic Report of the President*, many macroeconomic data sets reported at monthly frequencies (at least for the most recent years) and those that display seasonal patterns are all seasonally adjusted. The major sources for macroeconomic time series, including *Citibase*, also seasonally adjust many of the series. Thus, the scope for using our own seasonal adjustment is often limited.

Sometimes, we do work with seasonally unadjusted data, and it is useful to know that simple methods are available for dealing with seasonality in regression models. Generally, we can include a

set of **seasonal dummy variables** to account for seasonality in the dependent variable, the independent variables, or both.

The approach is simple. Suppose that we have monthly data, and we think that seasonal patterns within a year are roughly constant across time. For example, because Christmas always comes at the same time of year, we can expect retail sales to be, on average, higher in months late in the year than in earlier months. Or, because weather patterns are broadly similar across years, housing starts in the Midwest will be higher on average during the summer months than the winter months. A general model for monthly data that captures these phenomena is

$$y_t = \beta_0 + \delta_1 feb_t + \delta_2 mar_t + \delta_3 apr_t + \cdots + \delta_{11} dec_t + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t, \qquad [10.41]$$

GOING FURTHER 10.5

In equation (10.41), what is the intercept for March? Explain why seasonal dummy variables satisfy the strict exogeneity assumption.

where feb_t, mar_t, . . . , dec_t are dummy variables indicating whether time period t corresponds to the appropriate month. In this formulation, January is the base month, and β_0 is the intercept for January. If there is no seasonality in y_t, once the x_{tj} have been controlled for, then δ_1 through δ_{11} are all zero. This is easily tested via an F test.

EXAMPLE 10.11 Effects of Antidumping Filings

In Example 10.5, we used monthly data (in the file BARIUM) that have not been seasonally adjusted. Therefore, we should add seasonal dummy variables to make sure none of the important conclusions change. It could be that the months just before the suit was filed are months where imports are higher or lower, on average, than in other months. When we add the 11 monthly dummy variables as in (10.41) and test their joint significance, we obtain p-value = .59, and so the seasonal dummies are jointly insignificant. In addition, nothing important changes in the estimates once statistical significance is taken into account. Krupp and Pollard (1996) actually used three dummy variables for the seasons (fall, spring, and summer, with winter as the base season), rather than a full set of monthly dummies; the outcome is essentially the same.

If the data are quarterly, then we would include dummy variables for three of the four quarters, with the omitted category being the base quarter. Sometimes, it is useful to interact seasonal dummies with some of the x_{tj} to allow the effect of x_{tj} on y_t to differ across the year.

Just as including a time trend in a regression has the interpretation of initially detrending the data, including seasonal dummies in a regression can be interpreted as **deseasonalizing** the data. For concreteness, consider equation (10.41) with $k = 2$. The OLS slope coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ on x_1 and x_2 can be obtained as follows:

(i) Regress each of y_t, x_{t1}, and x_{t2} on a constant and the monthly dummies, feb_t, mar_t, . . . , dec_t, and save the residuals, say, $\ddot{y}_t$, $\ddot{x}_{t1}$, and $\ddot{x}_{t2}$, for all $t = 1, 2, \ldots, n$. For example,

$$\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 feb_t - \hat{\alpha}_2 mar_t - \cdots - \hat{\alpha}_{11} dec_t.$$

This is one method of deseasonalizing a monthly time series. A similar interpretation holds for $\ddot{x}_{t1}$ and $\ddot{x}_{t2}$.

(ii) Run the regression, without the monthly dummies, of $\ddot{y}_t$ on $\ddot{x}_{t1}$ and $\ddot{x}_{t2}$ [just as in (10.37)]. This gives $\hat{\beta}_1$ and $\hat{\beta}_2$.

In some cases, if y_t has pronounced seasonality, a better goodness-of-fit measure is an R-squared based on the deseasonalized y_t. This nets out any seasonal effects that are not explained by the x_{tj}.

Wooldridge (1991a) suggests specific degrees-of-freedom adjustments, or one may simply use the adjusted *R*-squared where the dependent variable has been deseasonalized.

Time series exhibiting seasonal patterns can be trending as well, in which case we should estimate a regression model with a time trend and seasonal dummy variables. The regressions can then be interpreted as regressions using both detrended and deseasonalized series. Goodness-of-fit statistics are discussed in Wooldridge (1991a): essentially, we detrend and deseasonalize y_t by regressing on both a time trend and seasonal dummies before computing *R*-squared or adjusted *R*-squared.

Summary

In this chapter, we have covered basic regression analysis with time series data. Under assumptions that parallel those for cross-sectional analysis, OLS is unbiased (under TS.1 through TS.3), OLS is BLUE (under TS.1 through TS.5), and the usual OLS standard errors, *t* statistics, and *F* statistics can be used for statistical inference (under TS.1 through TS.6). Because of the temporal correlation in most time series data, we must explicitly make assumptions about how the errors are related to the explanatory variables in all time periods and about the temporal correlation in the errors themselves. The classical linear model assumptions can be pretty restrictive for time series applications, but they are a natural starting point. We have applied them to both static regression and finite distributed lag models.

Logarithms and dummy variables are used regularly in time series applications and in event studies. We also discussed index numbers and time series measured in terms of nominal and real dollars.

Trends and seasonality can be easily handled in a multiple regression framework by including time and seasonal dummy variables in our regression equations. We presented problems with the usual *R*-squared as a goodness-of-fit measure and suggested some simple alternatives based on detrending or deseasonalizing.

CLASSICAL LINEAR MODEL ASSUMPTIONS FOR TIME SERIES REGRESSION

Following is a summary of the six classical linear model (CLM) assumptions for time series regression applications. Assumptions TS.1 through TS.5 are the time series versions of the Gauss-Markov assumptions (which implies that OLS is BLUE and has the usual sampling variances). We only needed TS.1, TS.2, and TS.3 to establish unbiasedness of OLS. As in the case of cross-sectional regression, the normality assumption, TS.6, was used so that we could perform exact statistical inference for any sample size.

Assumption TS.1 (Linear in Parameters)

The stochastic process $\{(x_{t1}, x_{t2}, \ldots, x_{tk}, y_t): t = 1, 2, \ldots, n\}$ follows the linear model

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t,$$

where $\{u_t: t = 1, 2, \ldots, n\}$ is the sequence of errors or disturbances. Here, n is the number of observations (time periods).

Assumption TS.2 (No Perfect Collinearity)

In the sample (and therefore in the underlying time series process), no independent variable is constant nor a perfect linear combination of the others.

Assumption TS.3 (Zero Conditional Mean)

For each t, the expected value of the error u_t, given the explanatory variables for *all* time periods, is zero. Mathematically, $\mathrm{E}(u_t|\mathbf{X}) = 0, t = 1, 2, \ldots, n$.

Assumption TS.3 replaces MLR.4 for cross-sectional regression, and it also means we do not have to make the random sampling assumption MLR.2. Remember, Assumption TS.3 implies that the error in each time period t is uncorrelated with all explanatory variables in *all* time periods (including, of course, time period t).

Assumption TS.4 (Homoskedasticity)

Conditional on $\mathbf{X}$, the variance of u_t is the same for all t: $\text{Var}(u_t|\mathbf{X}) = \text{Var}(u_t) = \sigma^2, t = 1, 2, \ldots, n$.

Assumption TS.5 (No Serial Correlation)

Conditional on $\mathbf{X}$, the errors in two different time periods are uncorrelated: $\text{Corr}(u_t, u_s|\mathbf{X}) = 0$, for all $t \neq s$.

Recall that we added the no serial correlation assumption, along with the homoskedasticity assumption, to obtain the same variance formulas that we derived for cross-sectional regression under random sampling. As we will see in Chapter 12, Assumption TS.5 is often violated in ways that can make the usual statistical inference very unreliable.

Assumption TS.6 (Normality)

The errors u_t are independent of $\mathbf{X}$ and are independently and identically distributed as $\text{Normal}(0,\sigma^2)$.

Key Terms

Autocorrelation
Base Period
Base Value
Contemporaneously Exogenous
Cumulative Effect
Deseasonalizing
Detrending
Event Study
Exponential Trend
Finite Distributed Lag (FDL) Model
Growth Rate
Impact Multiplier
Impact Propensity
Index Number
Lag Distribution
Linear Time Trend
Long-Run Elasticity
Long-Run Multiplier
Long-Run Propensity (LRP)
Seasonal Dummy Variables
Seasonality
Seasonally Adjusted
Serial Correlation
Short-Run Elasticity
Spurious Regression Problem
Static Model
Stochastic Process
Strictly Exogenous
Time Series Process
Time Trend

Problems

1 Decide if you agree or disagree with each of the following statements and give a brief explanation of your decision:

(i) Like cross-sectional observations, we can assume that most time series observations are independently distributed.

(ii) The OLS estimator in a time series regression is unbiased under the first three Gauss-Markov assumptions.

(iii) A trending variable cannot be used as the dependent variable in multiple regression analysis.

(iv) Seasonality is not an issue when using annual time series observations.

2 Let $gGDP_t$ denote the annual percentage change in gross domestic product and let int_t denote a short-term interest rate. Suppose that $gGDP_t$ is related to interest rates by

$$gGDP_t = \alpha_0 + \delta_0 int_t + \delta_1 int_{t-1} + u_t,$$

where u_t is uncorrelated with int_t, int_{t-1}, and all other past values of interest rates. Suppose that the Federal Reserve follows the policy rule:

$$int_t = \gamma_0 + \gamma_1(gGDP_{t-1} - 3) + v_t,$$

where $\gamma_1 > 0$. (When last year's GDP growth is above 3%, the Fed increases interest rates to prevent an "overheated" economy.) If v_t is uncorrelated with all past values of int_t and u_t, argue that int_t must be correlated with u_{t-1}. (*Hint*: Lag the first equation for one time period and substitute for $gGDP_{t-1}$ in the second equation.) Which Gauss-Markov assumption does this violate?

3 Suppose y_t follows a second order FDL model:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t.$$

Let z^* denote the *equilibrium value* of z_t and let y^* be the equilibrium value of y_t, such that

$$y^* = \alpha_0 + \delta_0 z^* + \delta_1 z^* + \delta_2 z^*.$$

Show that the change in y^*, due to a change in z^*, equals the long-run propensity times the change in z^*:

$$\Delta y^* = LRP \cdot \Delta z^*.$$

This gives an alternative way of interpreting the LRP.

4 When the three event indicators *befile6*, *affile6*, and *afdec6* are dropped from equation (10.22), we obtain $R^2 = .281$ and $\bar{R}^2 = .264$. Are the event indicators jointly significant at the 10% level?

5 Suppose you have quarterly data on new housing starts, interest rates, and real per capita income. Specify a model for housing starts that accounts for possible trends and seasonality in the variables.

6 In Example 10.4, we saw that our estimates of the individual lag coefficients in a distributed lag model were very imprecise. One way to alleviate the multicollinearity problem is to assume that the δ_j follow a relatively simple pattern. For concreteness, consider a model with four lags:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \delta_3 z_{t-3} + \delta_4 z_{t-4} + u_t.$$

Now, let us assume that the δ_j follow a quadratic in the lag, j:

$$\delta_j = \gamma_0 + \gamma_1 j + \gamma_2 j^2,$$

for parameters γ_0, γ_1, and γ_2. This is an example of a *polynomial distributed lag (PDL) model.*

(i) Plug the formula for each δ_j into the distributed lag model and write the model in terms of the parameters γ_h, for $h = 0, 1, 2$.

(ii) Explain the regression you would run to estimate the γ_h.

(iii) The polynomial distributed lag model is a restricted version of the general model. How many restrictions are imposed? How would you test these? (*Hint:* Think F test.)

7 In Example 10.4, we wrote the model that explicitly contains the long-run propensity, θ_0, as

$$gfr_t = \alpha_0 + \theta_0 pe_t + \delta_1(pe_{t-1} - pe_t) + \delta_2(pe_{t-2} - pe_t) + u,$$

where we omit the other explanatory variables for simplicity. As always with multiple regression analysis, θ_0 should have a ceteris paribus interpretation. Namely, if pe_t increases by one (dollar) holding $(pe_{t-1} - pe_t)$ and $(pe_{t-2} - pe_t)$ fixed, gfr_t should change by θ_0.

(i) If $(pe_{t-1} - pe_t)$ and $(pe_{t-2} - pe_t)$ are held fixed but pe_t is increasing, what must be true about changes in pe_{t-1} and pe_{t-2}?

(ii) How does your answer in part (i) help you to interpret θ_0 in the above equation as the LRP?

8 In the linear model given in equation (10.8), the explanatory variables $\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})$ are said to be *sequentially exogenous* (sometimes called *weakly exogenous*) if

$$\mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1) = 0, t = 1, 2, \ldots,$$

so that the errors are unpredictable given current and all *past* values of the explanatory variables.

(i) Explain why sequential exogeneity is implied by strict exogeneity.

(ii) Explain why contemporaneous exogeneity is implied by sequential exogeneity.

(iii) Are the OLS estimators generally unbiased under the sequential exogeneity assumption? Explain.

(iv) Consider a model to explain the annual rate of HIV infections (*HIVrate*) as a distributed lag of per capita condom usage (*pccon*) for a state, region, or province:

$$\mathrm{E}(HIVrate_t|pccon_t, pccont_{t-1}, \ldots,) = \alpha_0 + \delta_0 pccon_t + \delta_1 pccon_{t-1} + \delta_2 pccon_{t-2} + \delta_3 pccon_{t-3}.$$

Explain why this model satisfies the sequential exogeneity assumption. Does it seem likely that strict exogeneity holds too?

Computer Exercises

C1 In October 1979, the Federal Reserve changed its policy of using finely tuned interest rate adjustments and instead began targeting the money supply. Using the data in INTDEF, define a dummy variable equal to 1 for years after 1979. Include this dummy in equation (10.15) to see if there is a shift in the interest rate equation after 1979. What do you conclude?

C2 Use the data in BARIUM for this exercise.

(i) Add a linear time trend to equation (10.22). Are any variables, other than the trend, statistically significant?

(ii) In the equation estimated in part (i), test for joint significance of all variables except the time trend. What do you conclude?

(iii) Add monthly dummy variables to this equation and test for seasonality. Does including the monthly dummies change any other estimates or their standard errors in important ways?

C3 Add the variable log(*prgnp*) to the minimum wage equation in (10.38). Is this variable significant? Interpret the coefficient. How does adding log(*prgnp*) affect the estimated minimum wage effect?

C4 Use the data in FERTIL3 to verify that the standard error for the LRP in equation (10.19) is about .030.

C5 Use the data in EZANDERS for this exercise. The data are on monthly unemployment claims in Anderson Township in Indiana, from January 1980 through November 1988. In 1984, an enterprise zone (EZ) was located in Anderson (as well as other cities in Indiana). [See Papke (1994) for details.]

(i) Regress log(*uclms*) on a linear time trend and 11 monthly dummy variables. What was the overall trend in unemployment claims over this period? (Interpret the coefficient on the time trend.) Is there evidence of seasonality in unemployment claims?

(ii) Add *ez*, a dummy variable equal to one in the months Anderson had an EZ, to the regression in part (i). Does having the enterprise zone seem to decrease unemployment claims? By how much? [You should use formula (7.10) from Chapter 7.]

(iii) What assumptions do you need to make to attribute the effect in part (ii) to the creation of an EZ?

C6 Use the data in FERTIL3 for this exercise.

(i) Regress gfr_t on t and t^2 and save the residuals. This gives a detrended gfr_t, say, $\ddot{gf}_t$.

(ii) Regress $\ddot{gf}_t$ on all of the variables in equation (10.35), including t and t^2. Compare the R-squared with that from (10.35). What do you conclude?

(iii) Reestimate equation (10.35) but add t^3 to the equation. Is this additional term statistically significant?

C7 Use the data set CONSUMP for this exercise.

(i) Estimate a simple regression model relating the growth in real per capita consumption (of nondurables and services) to the growth in real per capita disposable income. Use the change in the logarithms in both cases. Report the results in the usual form. Interpret the equation and discuss statistical significance.

(ii) Add a lag of the growth in real per capita disposable income to the equation from part (i). What do you conclude about adjustment lags in consumption growth?

(iii) Add the real interest rate to the equation in part (i). Does it affect consumption growth?

C8 Use the data in FERTIL3 for this exercise.

(i) Add pe_{t-3} and pe_{t-4} to equation (10.19). Test for joint significance of these lags.

(ii) Find the estimated long-run propensity and its standard error in the model from part (i). Compare these with those obtained from equation (10.19).

(iii) Estimate the polynomial distributed lag model from Problem 6. Find the estimated LRP and compare this with what is obtained from the unrestricted model.

C9 Use the data in VOLAT for this exercise. The variable *rsp500* is the monthly return on the Standard & Poor's 500 stock market index, at an annual rate. (This includes price changes as well as dividends.) The variable *i3* is the return on three-month T-bills, and *pcip* is the percentage change in industrial production; these are also at an annual rate.

(i) Consider the equation

$$rsp500_t = \beta_0 + \beta_1 pcip_t + \beta_2 i3_t + u_t.$$

What signs do you think β_1 and β_2 should have?

(ii) Estimate the previous equation by OLS, reporting the results in standard form. Interpret the signs and magnitudes of the coefficients.

(iii) Which of the variables is statistically significant?

(iv) Does your finding from part (iii) imply that the return on the S&P 500 is predictable? Explain.

C10 Consider the model estimated in (10.15); use the data in INTDEF.

(i) Find the correlation between *inf* and *def* over this sample period and comment.

(ii) Add a single lag of *inf* and *def* to the equation and report the results in the usual form.

(iii) Compare the estimated LRP for the effect of inflation with that in equation (10.15). Are they vastly different?

(iv) Are the two lags in the model jointly significant at the 5% level?

C11 The file TRAFFIC2 contains 108 monthly observations on automobile accidents, traffic laws, and some other variables for California from January 1981 through December 1989. Use this data set to answer the following questions.

(i) During what month and year did California's seat belt law take effect? When did the highway speed limit increase to 65 miles per hour?

(ii) Regress the variable log(*totacc*) on a linear time trend and 11 monthly dummy variables, using January as the base month. Interpret the coefficient estimate on the time trend. Would you say there is seasonality in total accidents?

(iii) Add to the regression from part (ii) the variables *wkends*, *unem*, *spdlaw*, and *beltlaw*. Discuss the coefficient on the unemployment variable. Does its sign and magnitude make sense to you?

(iv) In the regression from part (iii), interpret the coefficients on *spdlaw* and *beltlaw*. Are the estimated effects what you expected? Explain.

(v) The variable *prcfat* is the percentage of accidents resulting in at least one fatality. Note that this variable is a percentage, not a proportion. What is the average of *prcfat* over this period? Does the magnitude seem about right?

(vi) Run the regression in part (iii) but use *prcfat* as the dependent variable in place of log(*totacc*). Discuss the estimated effects and significance of the speed and seat belt law variables.

C12 (i) Estimate equation (10.2) using all observations in PHILLIPS and report the results in the usual form. How many observations do you have now?

(ii) Compare the estimates from part (i) with those in equation (10.14). In particular, does adding the extra years help in obtaining an estimated tradeoff between inflation and unemployment? Explain.

(iii) Now run the regression using only the years 2007 through 2017. How do these estimates differ from those in equation (10.14)? Are the estimates using the most recent seven years precise enough to draw any firm conclusions? Explain.

(iv) Consider a simple regression setup in which we start with n time series observations and then split them into an early time period and a later time period. In the first time period we have n_1 observations and in the second period n_2 observations. Draw on the previous parts of this exercise to evaluate the following statement: "Generally, we can expect the slope estimate using all n observations to be roughly equal to a weighted average of the slope estimates on the early and later subsamples, where the weights are n_1/n and n_2/n, respectively."

C13 Use the data in MINWAGE for this exercise. In particular, use the employment and wage series for sector 232 (Men's and Boys' Furnishings). The variable *gwage232* is the monthly growth (change in logs) in the average wage in sector 232, *gemp232* is the growth in employment in sector 232, *gmwage* is the growth in the federal minimum wage, and *gcpi* is the growth in the (urban) Consumer Price Index.

(i) Run the regression *gwage232* on *gmwage*, *gcpi*. Do the sign and magnitude of $\hat{\beta}_{gmwage}$ make sense to you? Explain. Is *gmwage* statistically significant?

(ii) Add lags 1 through 12 of *gmwage* to the equation in part (i). Do you think it is necessary to include these lags to estimate the long-run effect of minimum wage growth on wage growth in sector 232? Explain.

(iii) Run the regression *gemp232* on *gmwage*, *gcpi*. Does minimum wage growth appear to have a contemporaneous effect on *gemp232*?

(iv) Add lags 1 through 12 to the employment growth equation. Does growth in the minimum wage have a statistically significant effect on employment growth, either in the short run or long run? Explain.

C14 Use the data in APPROVAL to answer the following questions. The data set consists of 78 months of data during the presidency of George W. Bush. (The data end in July 2007, before Bush left office.) In addition to economic variables and binary indicators of various events, it includes an approval rate, *approve*, collected by Gallup. (Caution: One should also attempt Computer Exercise C14 in Chapter 11 to gain a more complete understanding of the econometric issues involved in analyzing these data.)

(i) What is the range of the variable *approve*? What is its average value?

(ii) Estimate the model

$$approve_t = \beta_0 + \beta_1 lcpifood_t + \beta_2 lrgasprice_t + \beta_3 unemploy_t + u_t,$$

where the first two variables are in logarithmic form, and report the estimates in the usual way.

(iii) Interpret the coefficients in the estimates from part (ii). Comment on the signs and sizes of the effects, as well as statistical significance.

(iv) Add the binary variables *sep*11 and *iraqinvade* to the equation from part (ii). Interpret the coefficients on the dummy variables. Are they statistically significant?

(v) Does adding the dummy variables in part (iv) change the other estimates much? Are any of the coefficients in part (iv) hard to rationalize?

(vi) Add *lsp*500 to the regression in part (iv). Controlling for other factors, does the stock market have an important effect on the presidential approval rating?

PART 3

Advanced Topics

We now turn to some more specialized topics that are not usually covered in a one-term, introductory course. Some of these topics require few more mathematical skills than the multiple regression analysis did in Parts 1 and 2. In Chapter 13, we show how to apply multiple regression to independently pooled cross sections. The issues raised are very similar to standard cross-sectional analysis, except that we can study how relationships change over time by including time dummy variables. Pooled cross sections can be used very effectively for policy analysis, where a policy is assigned at a group level and we have not only at least one control group, but also periods before and after the intervention. We also illustrate how panel data sets can be analyzed in a regression framework. Chapter 14 covers more advanced panel data methods that are nevertheless used routinely in applied work.

Chapters 15 and 16 investigate the problem of endogenous explanatory variables. In Chapter 15, we introduce the method of instrumental variables as a way of solving the omitted variable problem as well as the measurement error problem. The method of two-stage least squares is used quite often in empirical economics and is indispensable for estimating simultaneous equation models, a topic we turn to in Chapter 16.

Chapter 17 covers some fairly advanced topics that are typically used in cross-sectional analysis, including models for limited dependent variables and methods for correcting sample selection bias. Chapter 18 heads in a different direction by covering some recent advances in time series econometrics that have proven to be useful in estimating dynamic relationships.

Chapter 19 should be helpful to students who must write either a term paper or some other paper in the applied social sciences. The chapter offers suggestions for how to select a topic, collect and analyze the data, and write the paper.

CHAPTER 13

Pooling Cross Sections across Time: Simple Panel Data Methods

Until now, we have covered multiple regression analysis using pure cross-sectional or pure time series data. Although these two cases arise often in applications, data sets that have both cross-sectional and time series dimensions are being used more and more often in empirical research. Multiple regression methods can still be used on such data sets. In fact, data with cross-sectional and time series aspects can often shed light on important policy questions. We will see several examples in this chapter.

We will analyze two kinds of data sets in this chapter. An **independently pooled cross section** is obtained by sampling randomly from a large population at different points in time (usually, but not necessarily, different years). For instance, in each year, we can draw a random sample on hourly wages, education, experience, and so on, from the population of working people in the United States. Or, in every other year, we draw a random sample on the selling price, square footage, number of bathrooms, and so on, of houses sold in a particular metropolitan area. From a statistical standpoint, these data sets have an important feature: they consist of *independently* sampled observations. This was also a key aspect in our analysis of cross-sectional data: among other things, it rules out correlation in the error terms across different observations.

An independently pooled cross section differs from a single random sample in that sampling from the population at different points in time likely leads to observations that are not identically distributed. For example, distributions of wages and education have changed over time in most countries. As we will see, this is easy to deal with in practice by allowing the intercept in a multiple regression

model, and in some cases the slopes, to change over time. We cover such models in Section 13-1. In Section 13-1, we discuss how pooling cross sections over time can be used to evaluate policy changes.

A **panel data** set, while having both a cross-sectional and a time series dimension, differs in some important respects from an independently pooled cross section. To collect panel data—sometimes called **longitudinal data**—we follow (or attempt to follow) the *same* individuals, families, firms, cities, states, or whatever, across time. For example, a panel data set on individual wages, hours, education, and other factors is collected by randomly selecting people from a population at a given point in time. Then, these *same* people are reinterviewed at several subsequent points in time. This gives us data on wages, hours, education, and so on, for the same group of people in different years.

Panel data sets are fairly easy to collect for school districts, cities, counties, states, and countries, and policy analysis is greatly enhanced by using panel data sets; we will see some examples in the following discussion. For the econometric analysis of panel data, we cannot assume that the observations are independently distributed across time. For example, unobserved factors (such as ability) that affect someone's wage in 1990 will also affect that person's wage in 1991; unobserved factors that affect a city's crime rate in 1985 will also affect that city's crime rate in 1990. For this reason, special models and methods have been developed to analyze panel data. In Sections 13-3, 13-4, and 13-5, we describe the straightforward method of differencing to remove time-constant, unobserved attributes of the units being studied. Because panel data methods are somewhat more advanced, we will rely mostly on intuition in describing the statistical properties of the estimation procedures, leaving detailed assumptions to the chapter appendix. We follow the same strategy in Chapter 14, which covers more complicated panel data methods.

13-1 Pooling Independent Cross Sections across Time

Many surveys of individuals, families, and firms are repeated at regular intervals, often each year. An example is the *Current Population Survey* (or CPS), which randomly samples households each year. (See, for example, CPS78_85, which contains data from the 1978 and 1985 CPS.) If a random sample is drawn at each time period, pooling the resulting random samples gives us an independently pooled cross section.

One reason for using independently pooled cross sections is to increase the sample size. By pooling random samples drawn from the same population, but at different points in time, we can get more precise estimators and test statistics with more power. Pooling is helpful in this regard only insofar as the relationship between the dependent variable and at least some of the independent variables remain constant over time.

As mentioned in the introduction, using pooled cross sections raises only minor statistical complications. Typically, to reflect the fact that the population may have different distributions in different time periods, we allow the intercept to differ across periods, usually years. This is easily accomplished by including dummy variables for all but one year, where the earliest year in the sample is usually chosen as the base year. It is also possible that the error variance changes over time, something we discuss later.

Sometimes, the pattern of coefficients on the year dummy variables is itself of interest. For example, a demographer may be interested in the following question: *After* controlling for education, has the pattern of fertility among women over age 35 changed between 1972 and 1984? The following example illustrates how this question is simply answered by using multiple regression analysis with **year dummy variables**.

EXAMPLE 13.1 Women's Fertility over Time

The data set in FERTIL1, which is similar to that used by Sander (1992), comes from the National Opinion Research Center's *General Social Survey* for the even years from 1972 to 1984, inclusively. We use these data to estimate a model explaining the total number of kids born to a woman (*kids*).

One question of interest is: After controlling for other observable factors, what has happened to fertility rates over time? The factors we control for are years of education, age, race, region of the country where living at age 16, and living environment at age 16. The estimates are given in Table 13.1.

The base year is 1972. The coefficients on the year dummy variables show a sharp drop in fertility in the early 1980s. For example, the coefficient on *y82* implies that, holding education, age, and other factors fixed, a woman had on average .52 less children, or about one-half a child, in 1982 than in 1972. This is a very large drop: holding *educ*, *age*, and the other factors fixed, 100 women in 1982 are predicted to have about 52 fewer children than 100 comparable women in 1972. Because we are controlling for education, this drop is separate from the decline in fertility that is due to the increase in average education levels. (The average years of education are 12.2 for 1972 and 13.3 for 1984.) The coefficients on *y82* and *y84* represent drops in fertility for reasons that are not captured in the explanatory variables.

Given that the 1982 and 1984 year dummies are individually quite significant, it is not surprising that as a group the year dummies are jointly very significant: the R-squared for the regression without the year dummies is .1019, and this leads to $F_{6,1111} = 5.87$ and p-value ≈ 0.

TABLE 13.1 Determinants of Women's Fertility

Dependent Variable: *kids*

Independent Variables	Coefficients	Standard Errors
educ	−.128	.018
age	.532	.138
age^2	−.0058	.0016
black	1.076	.174
east	.217	.133
northcen	.363	.121
west	.198	.167
farm	−.053	.147
othrural	−.163	.175
town	.084	.124
smcity	.212	.160
y74	.268	.173
y76	−.097	.179
y78	−.069	.182
y80	−.071	.183
y82	−.522	.172
y84	−.545	.175
constant	−7.742	3.052
$n = 1{,}129$ $R^2 = .1295$ $\bar{R}^2 = .1162$		

Women with more education have fewer children, and the estimate is very statistically significant. Other things being equal, 100 women with a college education will have about 51 fewer children on average than 100 women with only a high school education: .128(4) = .512. Age has a diminishing effect on fertility. (The turning point in the quadratic is at about *age* = 46, by which time most women have finished having children.)

The model estimated in Table 13.1 assumes that the effect of each explanatory variable, particularly education, has remained constant. This may or may not be true; you will be asked to explore this issue in Computer Exercise C1.

Finally, there may be heteroskedasticity in the error term underlying the estimated equation. This can be dealt with using the methods in Chapter 8. There is one interesting difference here: now, the error variance may change over time even if it does not change with the values of *educ*, *age*, *black*, and so on. The heteroskedasticity-robust standard errors and test statistics are nevertheless valid. The Breusch-Pagan test would be obtained by regressing the squared OLS residuals on *all* of the independent variables in Table 13.1, including the year dummies. (For the special case of the White statistic, the fitted values $\widehat{kids}$ and the squared fitted values are used as the independent variables, as always.) A weighted least squares procedure should account for variances that possibly change over time. In the procedure discussed in Section 8-4, year dummies would be included in equation (8.32).

GOING FURTHER 13.1

In reading Table 13.1, someone claims that, if everything else is equal in the table, a black woman is expected to have one more child than a nonblack woman. Do you agree with this claim?

We can also interact a year dummy variable with key explanatory variables to see if the effect of that variable has changed over a certain time period. The next example examines how the return to education and the gender gap have changed from 1978 to 1985.

EXAMPLE 13.2 **Changes in the Return to Education and the Gender Wage Gap**

A log(*wage*) equation (where *wage* is hourly wage) pooled across the years 1978 (the base year) and 1985 is

$$\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85 \cdot educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_5 y85 \cdot female + u, \quad [13.1]$$

where most explanatory variables should by now be familiar. The variable *union* is a dummy variable equal to one if the person belongs to a union, and zero otherwise. The variable *y85* is a dummy variable equal to one if the observation comes from 1985 and zero if it comes from 1978. There are 550 people in the sample in 1978 and a different set of 534 people in 1985.

The intercept for 1978 is β_0, and the intercept for 1985 is $\beta_0 + \delta_0$. The return to education in 1978 is β_1, and the return to education in 1985 is $\beta_1 + \delta_1$. Therefore, δ_1 measures how the return to another year of education has changed over the seven-year period. Finally, in 1978, the log(*wage*) differential between women and men is β_5; the differential in 1985 is $\beta_5 + \delta_5$. Thus, we can test the null hypothesis that nothing has happened to the gender differential over this seven-year period by testing H_0: $\delta_5 = 0$. The alternative that the gender differential has been *reduced* is H_1: $\delta_5 < 0$. For simplicity, we have assumed that experience and union membership have the same effect on wages in both time periods.

Before we present the estimates, there is one other issue we need to address—namely, hourly wage here is in nominal (or current) dollars. Because nominal wages grow simply due to inflation, we are really interested in the effect of each explanatory variable on real wages. Suppose that we settle on measuring wages in 1978 dollars. This requires deflating 1985 wages to 1978 dollars. (Using

the Consumer Price Index for the 1997 *Economic Report of the President*, the deflation factor is 107.6/65.2 ≈ 1.65.) Although we can easily divide each 1985 wage by 1.65, it turns out that this is not necessary, *provided* a 1985 year dummy is included in the regression *and* log(*wage*) (as opposed to *wage*) is used as the dependent variable. Using real or nominal wage in a logarithmic functional form only affects the coefficient on the year dummy, *y85*. To see this, let *P85* denote the deflation factor for 1985 wages (1.65, if we use the CPI). Then, the log of the real wage for each person *i* in the 1985 sample is

$$\log(wage_i/P85) = \log(wage_i) - \log(P85).$$

Now, while $wage_i$ differs across people, *P85* does not. Therefore, log(*P85*) will be absorbed into the intercept for 1985. (This conclusion would change if, for example, we used a different price index for people living in different parts of the country.) The bottom line is that, for studying how the return to education or the gender gap has changed, we do not need to turn nominal wages into real wages in equation (13.1). Computer Exercise C2 asks you to verify this for the current example.

If we forget to allow different intercepts in 1978 and 1985, the use of nominal wages can produce seriously misleading results. If we use *wage* rather than log(*wage*) as the dependent variable, it is important to use the real wage and to include a year dummy.

The previous discussion generally holds when using dollar values for either the dependent or independent variables. Provided the dollar amounts appear in logarithmic form and dummy variables are used for all time periods (except, of course, the base period), the use of aggregate price deflators will only affect the intercepts; none of the slope estimates will change.

Now, we use the data in CPS78_85 to estimate the equation:

$$\begin{aligned}\log(wage) = {} & .459 + .118\,y85 + .0747\,educ + .0185\,y85{\cdot}educ \\ & \quad (.093)\ \ (.124)\ \ \ \ (.0067)\ \ \ \ (.0094) \\ & + .0296\,exper - .00040\,exper^2 + .202\,union \\ & \quad (.0036)\ \ \ \ (.00008)\ \ \ \ (.030) \\ & - .317\,female + .085\,y85{\cdot}female \\ & \quad (.037)\ \ \ \ (.051) \\ n = {} & 1{,}084;\ R^2 = .426;\ \overline{R}^2 = .422.\end{aligned} \qquad [13.2]$$

The return to education in 1978 is estimated to be about 7.5%; the return to education in 1985 is about 1.85 percentage points *higher*, or about 9.35%. Because the *t* statistic on the interaction term is .0185/.0094 ≈ 1.97, the difference in the return to education is statistically significant at the 5% level against a two-sided alternative.

What about the gender gap? In 1978, other things being equal, a woman earned about 31.7% less than a man (27.2% is the more accurate estimate). In 1985, the gap in log(*wage*) is −.317 + .085 = −.232. Therefore, the gender gap appears to have fallen from 1978 to 1985 by about 8.5 percentage points. The *t* statistic on the interaction term is about 1.67, which means it is significant at the 5% level against the positive one-sided alternative.

What happens if we interact *all* independent variables with *y85* in equation (13.2)? This is identical to estimating two separate equations, one for 1978 and one for 1985. Sometimes, this is desirable. For example, in Chapter 7, we discussed a study by Krueger (1993), in which he estimated the return to using a computer on the job. Krueger estimates two separate equations, one using the 1984 CPS and the other using the 1989 CPS. By comparing how the return to education changes across time and whether or not computer usage is controlled for, he estimates that one-third to one-half of the observed increase in the return to education over the five-year period can be attributed to increased computer usage. [See Tables VIII and IX in Krueger (1993).]

13-1a The Chow Test for Structural Change across Time

In Chapter 7, we discussed how the Chow test—which is simply an F test—can be used to determine whether a multiple regression function differs across two groups. We can apply that test to two different time periods as well. One form of the test obtains the sum of squared residuals from the pooled estimation as the restricted SSR. The unrestricted SSR is the sum of the SSRs for the two separately estimated time periods. The mechanics of computing the statistic are exactly as they were in Section 7-4. A heteroskedasticity-robust version is also available (see Section 8-2).

Example 13.2 suggests another way to compute the Chow test for two time periods by interacting each variable with a year dummy for one of the two years and testing for joint significance of the year dummy and all of the interaction terms. Because the intercept in a regression model often changes over time (due to, say, inflation in the housing price example), this full-blown Chow test can detect such changes. It is usually more interesting to allow for an intercept difference and then to test whether certain slope coefficients change over time (as we did in Example 13.2).

A Chow test can also be computed for more than two time periods. Just as in the two-period case, it is usually more interesting to allow the intercepts to change over time and then test whether the slope coefficients have changed over time. We can test the constancy of slope coefficients generally by interacting all of the time-period dummies (except that defining the base group) with one, several, or all of the explanatory variables and test the joint significance of the interaction terms. Computer Exercises C1 and C2 are examples. For many time periods and explanatory variables, constructing a full set of interactions can be tedious. Alternatively, we can adapt the approach described in part (vi) of Computer Exercise C11 in Chapter 7. First, estimate the restricted model by doing a pooled regression allowing for different time intercepts; this gives SSR_r. Then, run a regression for each of the, say, T time periods and obtain the sum of squared residuals for each time period. The unrestricted sum of squared residuals is obtained as $SSR_{ur} = SSR_1 + SSR_2 + \cdots + SSR_T$. If there are k explanatory variables (not including the intercept or the time dummies) with T time periods, then we are testing $(T - 1)k$ restrictions, and there are $T + Tk$ parameters estimated in the unrestricted model. So, if $n = n_1 + n_2 + \cdots + n_T$ is the total number of observations, then the df of the F test are $(T - 1)k$ and $n - T - Tk$. We compute the F statistic as usual: $[(SSR_r - SSR_{ur})/SSR_{ur}][(n - T - Tk)/(T - 1)k]$. Unfortunately, as with any F test based on sums of squared residuals or R-squareds, this test is not robust to heteroskedasticity (including changing variances across time). To obtain a heteroskedasticity-robust test, we must construct the interaction terms and do a pooled regression.

13-2 Policy Analysis with Pooled Cross Sections

Pooled cross sections can be very useful for evaluating the impact of a certain event or policy. The following example of an event study shows how two cross-sectional data sets, collected before and after the occurrence of an event, can be used to determine the effect on economic outcomes.

EXAMPLE 13.3 **Effect of a Garbage Incinerator's Location on Housing Prices**

Kiel and McClain (1995) studied the effect that a new garbage incinerator had on housing values in North Andover, Massachusetts. They used many years of data and a fairly complicated econometric analysis. We will use two years of data and some simplified models, but our analysis is similar.

The rumor that a new incinerator would be built in North Andover began after 1978, and construction began in 1981. The incinerator was expected to be in operation soon after the start of construction; the incinerator actually began operating in 1985. We will use data on prices of houses that sold in 1978 and another sample on those that sold in 1981. The hypothesis is that the price of houses located near the incinerator would fall relative to the price of more distant houses.

For illustration, we define a house to be near the incinerator if it is within three miles. [In Computer Exercise C3, you are instead asked to use the actual distance from the house to the incinerator, as in Kiel and McClain (1995).] We will start by looking at the dollar effect on housing prices. This requires us to measure price in constant dollars. We measure all housing prices in 1978 dollars, using the Boston housing price index. Let *rprice* denote the house price in real terms.

A naive analyst would use only the 1981 data and estimate a very simple model:

$$rprice = \gamma_0 + \gamma_1 nearinc + u, \qquad [13.3]$$

where *nearinc* is a binary variable equal to one if the house is near the incinerator, and zero otherwise. Estimating this equation using the data in KIELMC gives

$$\begin{aligned} \widehat{rprice} &= 101{,}307.5 - 30{,}688.27\, nearinc \\ &\quad (3{,}093.0) \quad (5{,}827.71) \\ n &= 142,\ R^2 = .165. \end{aligned} \qquad [13.4]$$

Because this is a simple regression on a single dummy variable, the intercept is the average selling price for homes not near the incinerator, and the coefficient on *nearinc* is the difference in the average selling price between homes near the incinerator and those that are not. The estimate shows that the average selling price for the former group was $30,688.27 less than for the latter group. The *t* statistic is greater than five in absolute value, so we can strongly reject the hypothesis that the average value for homes near and far from the incinerator are the same.

Unfortunately, equation (13.4) does *not* imply that the siting of the incinerator is causing the lower housing values. In fact, if we run the same regression for 1978 (before the incinerator was even rumored), we obtain

$$\begin{aligned} \widehat{rprice} &= 82{,}517.23 - 18{,}824.37\, nearinc \\ &\quad (2{,}653.79) \quad (4{,}744.59) \\ n &= 179,\ R^2 = .082. \end{aligned} \qquad [13.5]$$

Therefore, even *before* there was any talk of an incinerator, the average value of a home near the site was $18,824.37 less than the average value of a home not near the site ($82,517.23); the difference is statistically significant, as well. This is consistent with the view that the incinerator was built in an area with lower housing values.

How, then, can we tell whether building a new incinerator depresses housing values? The key is to look at how the coefficient on *nearinc* changed between 1978 and 1981. The difference in average housing value was much larger in 1981 than in 1978 ($30,688.27 versus $18,824.37), even as a percentage of the average value of homes not near the incinerator site. The difference in the two coefficients on *nearinc* is

$$\hat{\delta}_1 = -30{,}688.27 - (-18{,}824.37) = -11{,}863.9.$$

This is our estimate of the effect of the incinerator on values of homes near the incinerator site. In empirical economics, $\hat{\delta}_1$ has become known as the **difference-in-differences** (**DD** or **DID**) **estimator** because it can be expressed as

$$\hat{\delta}_1 = (\overline{rprice}_{81,\,nr} - \overline{rprice}_{81,\,fr}) - (\overline{rprice}_{78,\,nr} - \overline{rprice}_{78,\,fr}), \qquad [13.6]$$

where *nr* stands for "near the incinerator site" and *fr* stands for "farther away from the site." In other words, $\hat{\delta}_1$ is the difference over time in the average difference of housing prices in the two locations.

To test whether $\hat{\delta}_1$ is statistically different from zero, we need to find its standard error by using a regression analysis. In fact, $\hat{\delta}_1$ can be obtained by estimating

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u, \qquad [13.7]$$

using the data pooled over both years. The intercept, β_0, is the average price of a home not near the incinerator in 1978. The parameter δ_0 captures changes in *all* housing values in North Andover from 1978 to 1981. [A comparison of equations (13.4) and (13.5) shows that housing values in North Andover, relative to the Boston housing price index, increased sharply over this period.] The coefficient on *nearinc*, β_1, measures the location effect that is *not* due to the presence of the incinerator: as we saw in equation (13.5), even in 1978, homes near the incinerator site sold for less than homes farther away from the site.

The parameter of interest is on the interaction term *y81·nearinc*: δ_1 measures the decline in housing values due to the new incinerator, provided we assume that houses both near and far from the site did not appreciate at different rates for other reasons.

The estimates of equation (13.7) are given in column (1) of Table 13.2. The only number we could not obtain from equations (13.4) and (13.5) is the standard error of $\hat{\delta}_1$. The t statistic on $\hat{\delta}_1$ is about -1.59, which is marginally significant against a one-sided alternative (p-value $\approx .057$).

Kiel and McClain (1995) included various housing characteristics in their analysis of the incinerator siting. There are two good reasons for doing this. First, the kinds of homes selling near the incinerator in 1981 might have been systematically different than those selling near the incinerator in 1978; if so, it can be important to control for such characteristics. Second, even if the relevant house characteristics did not change, including them can greatly reduce the error variance, which can then shrink the standard error of $\hat{\delta}_1$. (See Section 6-3 for discussion.) In column (2), we control for the age of the houses, using a quadratic. This substantially increases the R-squared (by reducing the residual variance). The coefficient on *y81·nearinc* is now much larger in magnitude, and its standard error is lower.

In addition to the age variables in column (2), column (3) controls for distance to the interstate in feet (*intst*), land area in feet (*land*), house area in feet (*area*), number of rooms (*rooms*), and number of baths (*baths*). This produces an estimate on *y81·nearinc* closer to that without any controls, but it yields a much smaller standard error: the t statistic for $\hat{\delta}_1$ is about -2.84. Therefore, we find a much more significant effect in column (3) than in column (1). The column (3) estimates are preferred because they control for the most factors and have the smallest standard errors (except in the constant, which is not important here). The fact that *nearinc* has a much smaller coefficient and is insignificant in column (3) indicates that the characteristics included in column (3) largely capture the housing characteristics that are most important for determining housing prices.

For the purpose of introducing the method, we used the level of real housing prices in Table 13.2. It makes more sense to use log(*price*) [or log(*rprice*)] in the analysis in order to get an approximate percentage effect. The basic model becomes

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u. \qquad [13.8]$$

TABLE 13.2 Effects of Incinerator Location on Housing Prices

Dependent Variable: *rprice*			
Independent Variable	**(1)**	**(2)**	**(3)**
constant	82,517.23 (2,726.91)	89,116.54 (2,406.05)	13,807.67 (11,166.59)
y81	18,790.29 (4,050.07)	21,321.04 (3,443.63)	13,928.48 (2,798.75)
nearinc	−18,824.37 (4,875.32)	9,397.94 (4,812.22)	3,780.34 (4,453.42)
y81·nearinc	−11,863.90 (7,456.65)	−21,920.27 (6,359.75)	−14,177.93 (4,987.27)
Other controls	No	*age*, age^2	Full Set
Observations	321	321	321
R-squared	.174	.414	.660

Now, $100 \cdot \delta_1$ is the approximate percentage reduction in housing value due to the incinerator. [Just as in Example 13.2, using log(*price*) versus log(*rprice*) only affects the coefficient on *y81*.] Using the same 321 pooled observations gives

$$\widehat{\log(price)} = 11.29 + .457\, y81 - .340\, nearinc - .063\, y81 \cdot nearinc \quad [13.9]$$
$$(.31)\ (.045) \qquad (.055) \qquad (.083)$$
$$n = 321,\ R^2 = .409.$$

The coefficient on the interaction term implies that, because of the new incinerator, houses near the incinerator lost about 6.3% in value. However, this estimate is not statistically different from zero. But when we use a full set of controls, as in column (3) of Table 13.2 (but with *intst*, *land*, and *area* appearing in logarithmic form), the coefficient on *y81*·*nearinc* becomes −.132 with a *t* statistic of about −2.53. Again, controlling for other factors turns out to be important. Using the logarithmic form, we estimate that houses near the incinerator were devalued by about 13.2%.

The methodology used in the previous example has numerous applications, especially when the data arise from a **natural experiment** (or a **quasi-experiment**). A natural experiment occurs when some exogenous event—often a change in government policy—changes the environment in which individuals, families, firms, or cities operate. A natural experiment always has a control group, which is not affected by the policy change, and a treatment group, which is thought to be affected by the policy change. Unlike a true experiment, in which treatment and control groups are randomly and explicitly chosen, the control and treatment groups in natural experiments arise from the particular policy change. To control for systematic differences between the control and treatment groups, we need two years of data, one before the policy change and one after the change. Thus, our sample is usefully broken down into four groups: the control group before the change, the control group after the change, the treatment group before the change, and the treatment group after the change.

Call *C* the control group and *T* the treatment group, letting *dT* equal unity for those in the treatment group *T*, and zero otherwise. Then, letting *d2* denote a dummy variable for the second (post-policy change) time period, the equation of interest is

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + other\ factors, \quad [13.10]$$

where *y* is the outcome variable of interest. As in Example 13.3, δ_1 measures the effect of the policy. Without other factors in the regression, $\hat{\delta}_1$ will be the difference-in-differences estimator:

$$\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{2,C}) - (\bar{y}_{1,T} - \bar{y}_{1,C}), \quad [13.11]$$

where the bar denotes average, the first subscript denotes the year, and the second subscript denotes the group. By simple rearrangement of (13.11), we can also write

$$\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{1,T}) - (\bar{y}_{2,C} - \bar{y}_{1,C}), \quad [13.12]$$

which provides a different interpretation of the DD estimator. The first term, $\bar{y}_{2,T} - \bar{y}_{1,T}$, is the difference in means over time for the treated group. This quantity would be a good estimator of the policy effect only if we can assume no external factors changed across the two time periods. (It is a before-after estimator applied to just the treated group.) To guard against this possibility, we compute the same trend in averages for the control group, $\bar{y}_{2,C} - \bar{y}_{1,C}$. By subtracting this from $\bar{y}_{2,T} - \bar{y}_{1,T}$ we hope to get a good estimator of the causal impact of the program or intervention.

The standard difference-in-differences setup is shown in Table 13.3. Like equations (13.11) and (13.12), Table 13.3 shows that the parameter δ_1 can be estimated in two ways: (1) Compute the differences in averages between the treatment and control groups in each time period, and then difference the results over time, as in equation (13.11); (2) Compute the change in averages over time for each of the treatment and control groups, and then difference these changes, as in equation (13.12). Naturally,

TABLE 13.3 Illustration of the Difference-in-Differences Estimator

	Before	After	After − Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment—Control	β_1	$\beta_1 + \delta_1$	δ_1

the estimate $\hat{\delta}_1$, which we can also be written as $\hat{\delta}_{DD}$, does not depend on how we do the differencing, but it is helpful to have the two different interpretations.

We can make a connection between the difference-in-differences framework and the potential outcomes framework that we discussed in previous chapters; see, for example, Section 7.6. The parameter δ_1 can be given an interpretation as an **average treatment effect (ATE)**, where the "treatment" is being in group T in the second time period.

When explanatory variables are added to equation (13.10) (to control for the fact that the populations sampled may differ systematically over the two periods), the OLS estimate of δ_1 no longer has the simple form of (13.11), but its interpretation is similar.

EXAMPLE 13.4 Effect of Worker Compensation Laws on Weeks out of Work

Meyer, Viscusi, and Durbin (1995) (hereafter, MVD) studied the length of time (in weeks) that an injured worker receives workers' compensation. On July 15, 1980, Kentucky raised the cap on weekly earnings that were covered by workers' compensation. An increase in the cap has no effect on the benefit for low-income workers, but it makes it less costly for a high-income worker to stay on workers' compensation. Therefore, the control group is low-income workers, and the treatment group is high-income workers; high-income workers are defined as those who were subject to the pre-policy change cap. Using random samples both before and after the policy change, MVD were able to test whether more generous workers' compensation causes people to stay out of work longer (everything else fixed). They started with a difference-in-differences analysis, using log(*durat*) as the dependent variable. Let *afchnge* be the dummy variable for observations after the policy change and *highearn* the dummy variable for high earners. Using the data in INJURY, the estimated equation, with standard errors in parentheses, is

$$\begin{aligned}\widehat{\log(durat)} &= 1.126 + .0077\, afchnge + .256\, highearn \\ &\qquad (0.031)\ \ (.0447)\qquad\quad (.047) \\ &\quad + .191\, afchnge{\cdot}highearn \\ &\qquad (.069) \\ n &= 5{,}626;\ R^2 = .021.\end{aligned} \qquad [13.13]$$

Therefore, $\hat{\delta}_1 = .191 (t = 2.77)$, which implies that the average length of time on workers' compensation for high earners increased by about 19% due to the increased earnings cap. The coefficient on *afchnge* is small and statistically insignificant: as is expected, the increase in the earnings cap has no effect on duration for low-income workers.

This is a good example of how we can get a fairly precise estimate of the effect of a policy change even though we cannot explain much of the variation in the dependent variable. The dummy variables in (13.13) explain only 2.1% of the variation in log(*durat*). This makes sense: there are clearly many factors, including severity of the injury, that affect how long someone receives workers' compensation. Fortunately, we have a very large sample size, and this allows us to get a significant t statistic.

MVD also added a variety of controls for gender, marital status, age, industry, and type of injury. This allows for the fact that the kinds of people and types of injuries may differ systematically by earnings group across the two years. Controlling for these factors turns out to have little effect on the estimate of δ_1. (See Computer Exercise C4.)

GOING FURTHER 13.2

What do you make of the coefficient and *t* statistic on *highearn* in equation (13.13)?

Sometimes, the two groups consist of people living in two neighboring states in the United States. For example, to assess the impact of changing cigarette taxes on cigarette consumption, we can obtain random samples from two states for two years. In State A, the control group, there was no change in the cigarette tax. In State B, the treatment group, the tax increased (or decreased) between the two years. The outcome variable would be a measure of cigarette consumption, and equation (13.10) can be estimated to determine the effect of the tax on cigarette consumption.

For an interesting survey on natural experiment methodology and several additional examples, see Meyer (1995).

13-2a Adding an Additional Control Group

One of the shortcomings of the traditional two-group, two-period difference-in-differences setup is that it assumes that any trends in the outcome, y, would trend at the same rate in the absence of the intervention. (This could be a positive or negative trend.) For example, suppose we are studying the effects of expanded health care for low-income families in a particular state. As in the Meyer, Viscusi, and Durbin (1995) application, we might use as a control group middle-income families that are not impacted by the policy change. In using the basic DD setup, we would have to assume that average health *trends* would be the same for the low-income and middle-income families in the absence of the intervention. This assumption is often known as the **parallel trends assumption**. Violation of the parallel trends assumption is a threat to the identification strategy used by DD, as can be seen by studying the expression for $\hat{\delta}_{DD}$ given in equation (13.12): the DD estimate is simply the difference in the estimated trends for the treatment and control groups.

One way to allow more flexibility is to collect information on a different control group. For example, suppose that in another state there was no intervention. If we think any differences in trends in health outcomes between low- and middle-income families is similar across states, we can include the state without the intervention as a control to obtain a more convincing estimate.

To be more concrete, let L denote low-income families and M middle-income families. Let B denote the state where the intervention occured and A the control state. Let dL be a dummy variable indicating low-income families, dB a dummy variable indicating state B, and $d2$ a dummy variable for the second time period. Now we estimate the equation

$$y = \beta_0 + \beta_1 dL + \beta_2 dB + \beta_3 dL \cdot dB + \delta_0 d2 + \delta_1 d2 \cdot dL + \delta_2 d2 \cdot dB + \delta_3 d2 \cdot dL \cdot dB + u, \qquad [13.14]$$

where y is some measure of health outcomes. Equation (13.14) includes each dummy variable separately, the three pairwise interactions, and the triple interaction term, $d2 \cdot dL \cdot dB$. This last term is the treatment indicator; the other terms act as controls that allow differences across time, income group, and state. Note that the trend for low-income people is allowed to be different from middle-income people through the term $d2 \cdot dL$, but in order to interpret δ_3 as the policy effect, we assume that any difference in trends between the L and M groups is the same across states in the absence of the intervention.

The easiest way to interpret the above equation is to study the OLS estimator of δ_3. After some tedious algebra, it can be shown that

$$\begin{aligned}\hat{\delta}_3 &= [(\bar{y}_{2,L,B} - \bar{y}_{1,L,B}) - (\bar{y}_{2,M,B} - \bar{y}_{1,M,B})] \\ &\quad - [(\bar{y}_{2,L,A} - \bar{y}_{1,L,A}) - (\bar{y}_{2,M,A} - \bar{y}_{1,M,A})] \\ &= \hat{\delta}_{DD,B} - \hat{\delta}_{DD,A} = \hat{\delta}_{DDD}.\end{aligned} \qquad [13.15]$$

The first term in brackets is the usual DD estimator using only the state that imposed the new policy. It uses as a control group middle-income families from the same state. The second term is the DD estimator in the state *not* imposing the new policy. If health trends between the L and M groups do not differ in state A, and there were no other intervention that would affect health outcomes, then $\hat{\delta}_{DD,A}$

should be roughly zero. In general, we estimate the policy effect by subtracting $\hat{\delta}_{DD,A}$ from $\hat{\delta}_{DD,B}$ as a way of accounting for possibly different trends in the L and M groups. If the differing trends in L and M also differ by state then even (13.15) will not produce a consistent estimator of the policy effect.

The estimator $\hat{\delta}_3$ is usually called the **difference-in-difference-in-differences (DDD) estimator**, and can be denoted $\hat{\delta}_{DDD}$. We obtain two DD estimators and then difference those. In obtaining the DDD estimator, it is convenient to use OLS applied to equation (13.14) because heteroskedasticity-robust standard errors are easy to obtain. Plus, as in the DD case, we can include control variables $x_1, \ldots, x_k$, either to account for compositional effects or to reduce the error variance in order to improve precision of $\hat{\delta}_{DDD}$.

13-2b A General Framework for Policy Analysis with Pooled Cross Sections

Another way to expand the basic DD methodology is to obtain multiple control and treatment groups as well as more than two time periods. We can create a very general framework for policy analysis by allowing a general pattern of interventions, where some units are never "treated" and others may be treated in different time periods. It is even possible that early in the study some units are subject to a policy but then later on the policy is dropped. As a word of warning, with general patterns of intervention it is a mistake to try to fit the problem in the basic DD or even DDD frameworks.

It is helpful to introduce an i subscript to represent an individual unit, which could be a person, a family, a firm, school, and so on. Each i belongs to a pair (g, t), where g is a group and t denotes a time period. Often the groups are based on geography, such as a city, county, state, or province, but we have already seen examples where the groups can be something else (low earners and high earners, for example). Most commonly, t represents a year, but it could be much shorter than a year or we could have time periods spread out more than a year apart.

In the general setting, we are interested in a policy intervention that applies at the group level. In order to be convincing, there should be a before-after period for at least some of the groups. Other groups may be control groups in that the policy is never implemented. In the simplest case, the policy is indicated by a dummy variable, say x_{gt}, which is one if group g in year t is subject to the policy intervention, and zero otherwise. It is very important to properly code this variable before undertaking any analysis. With many groups and time periods, the pattern of zeros and ones for x_{gt} can be complex. The complexity, which can result from policy interventions being staggered across groups, and policies being rescinded, adds power to our ability to estimate the effects of policy changes. It is important to not think that x_{gt} can always be constructed by interacting dummy variables indicating groups and time periods, as in the basic DD setup.

Given the policy variable x_{gt}, we can now write down an equation that can be used to estimate the policy effect. A flexible (but not completely flexible) model is

$$y_{igt} = \lambda_t + \alpha_g + \beta x_{gt} + \mathbf{z}_{igt}\gamma + u_{igt}, \; i = 1, \ldots, N_{gt} \qquad [13.16]$$
$$g = 1, \ldots, G; \; t = 1, \ldots, T,$$

where the notation shows that the group/time cell (g, t) has N_{gt} observations. The variable y_{igt} is measured at the unit level, as are the explanatory variables $\mathbf{z}_{igt}$. Recall that $\mathbf{z}_{igt}\gamma$ is shorthand for several explanatory variables multiplied by a coefficient.

The parameters λ_t are the aggregate time effects that capture external factors. For example, if g indexes states, the λ_t can be country-wide factors that affect all states equally. In most policy studies it is very important to account for such factors, as policy implementation tends to be bunched in different time periods. If we do not include the λ_t, then we may spuriously conclude that a policy had an impact. Or, we may estimate little or no policy effect where there is one. The group effects, α_g—for example, a state effect if g indexes states—account for systematic differences in groups that are constant across time. Policy implementation tends to depend on group characteristics that we may not be able to fully measure, and these same factors may influence y_{itg}. This is true whether g represents entities such as counties and states or whether they are, say, different age groups or income groups.

In estimating (13.16), we account for the time and group effects simply by including dummy variables. In other words, we define dummy variables, say dt, for each time period, and group dummy variables, dg, for each group. In practice, one includes an intercept and excludes one group and one time period (usually the first, but any one will do.) Then we estimate (13.16) using pooled OLS, where the pooling is across all individuals across all (g, t) pairs. The coefficient of interest is β. The variables $\mathbf{z}_{igt}$ can include measured variables that change only at the (g, t) level but also, as the i subscript indicates, individual-specific covariates. When we take the policy assignment as fixed and view estimation uncertainty through the sampling error, proper inference is obtained using heteroskedasticity-robust standard errors in the pooled OLS estimation.

The setup in equation (13.16) can be applied to important problems such as studying the labor market impacts of minimum wages. In the United States, minimum wages can vary at the city level, in which case g indexes county, although it is most common to study state level variation. The individual outcomes y_{igt} can be hourly wage (probably its log) or employment status. It could be very important to account for both time and city (or state) effects. In addition, we might have information on education, workforce experience, and background variables for individuals; these controls would be included in $\mathbf{z}_{igt}$.

Equation (3.16) imposes its own version of a parallel trend assumption because the effects λ_t have the same impact for all groups g. In particular, drop the variables $\mathbf{z}_{igt}$ and set x_{gt} to zero for all (g, t). Then, for a given group g, the mean value of y_{igt} is simply traced about by λ_t. One way to relax that assumption is to use **group-specific** linear time trends, at least if we have $T \geq 3$ time periods. In particular, we replace (13.16) with

$$y_{igt} = \lambda_t + \alpha_g + \psi_g t + \beta x_{gt} + \mathbf{z}_{igt}\gamma + u_{igt} \qquad [13.17]$$

where ψ_g captures the linear trend for group g. Notice that we still want aggregate effects included because the term $\psi_g t$ imposes linear trends on each group. We still want the λ_t to account for nonlinear aggregate time effects. In estimation, we will lose another λ_t because we have partly accounted for aggregate time effects with the group-specific trend.

In the minimum wage example, it is easy to imagine that minimum wages and labor market outcomes have trends that differ by city (or state).

Why should we use only linear group-specific trends? In fact, with lots of time periods we can include more complicated trends, such as a group-specific quadratic. But the more terms we include, the more variation in the policy indicator x_{gt} we require in order to pin down any effects of the policy. In the extreme case, one might think of including separate dummy variables for all (g, t) pairs, represented by, say,

$$y_{igt} = \theta_{gt} + \beta x_{gt} + \mathbf{z}_{igt}\gamma + u_{igt}, \qquad [13.18]$$

where θ_{gt} is a different intercept for each (g, t) pair. Such a formulation is more general than any of the previous equations, including (13.17). Unfortunately, it is also useless for estimating β because x_{gt} varies only at the (g, t) level, and is, therefore, perfectly collinear with the intercepts.

If interest lies in elements of γ—for example, the policy of interest applies to different units within at least some (g, t) pairs—then (13.18) becomes attractive. Operationally, rather than just including separate dummies for time and separate dummies for groups, as in (13.16), one includes a full set of time-group interactions: $dt \cdot dg$ for all $t = 1, \ldots, T$ and $g = 1, \ldots, G$. This allows each group to have its own very flexible time trend.

Extensions to more than one policy variable are straightforward, and the policy variables need not be binary indicators. For example, the vector $\mathbf{x}_{gt}$ can include, say, the state-level minimum wage for state g in year t, along with a dummy variable equal to unity if the state is a right-to-work state. Or, maybe we have individual-level health outcomes, y_{igt}, and $\mathbf{x}_{gt}$ is a vector (collection) of state-level health policy variables, which could be continuous or discrete. Then (13.16) becomes

$$\begin{aligned} y_{igt} &= \lambda_t + \alpha_g + \mathbf{x}_{gt}\boldsymbol{\beta} + \mathbf{z}_{igt}\boldsymbol{\gamma} + u_{igt} \qquad [13.19] \\ &= \lambda_t + \alpha_g + \beta_1 x_{gt1} + \cdots + \beta_K x_{gtK} + \gamma_1 z_{igt1} + \cdots + \gamma_J z_{igtJ} + u_{igt}. \end{aligned}$$

The β_j measure the (ceteris paribus) effect of different policies. If we think the policy takes time to have its full impact, we can even include a lagged policy indicator. For example, if p_{gt} is the policy indicator in time period t, we might estimate an equation such as

$$y_{igt} = \lambda_t + \alpha_g + \beta_1 p_{gt} + \beta_2 p_{g,t-1} + \beta_3 p_{g,t-2} + \gamma_1 z_{igt1} + \cdots + \gamma_J z_{igtJ} + u_{igt}.$$

Naturally, including lags requires more time periods. Generally, equation (13.17) gets modified in a similar way.

13-3 Two-Period Panel Data Analysis

We now turn to the analysis of the simplest kind of panel data: for a cross section of individuals, schools, firms, cities, or whatever, we have two years of data; call these $t = 1$ and $t = 2$. These years need not be adjacent, but $t = 1$ corresponds to the earlier year. For example, the file CRIME2 contains data on (among other things) crime and unemployment rates for 46 cities for 1982 and 1987. Therefore, $t = 1$ corresponds to 1982, and $t = 2$ corresponds to 1987.

What happens if we use the 1987 cross section and run a simple regression of *crmrte* on *unem*? We obtain

$$\widehat{crmrte} = 128.38 - 4.16\, unem$$
$$(20.76)\quad (3.42)$$
$$n = 46,\ R^2 = .033.$$

If we interpret the estimated equation causally, it implies that an increase in the unemployment rate *lowers* the crime rate. This is certainly not what we expect. The coefficient on *unem* is not statistically significant at standard significance levels: at best, we have found no link between crime and unemployment rates.

As we have emphasized throughout this text, this simple regression equation likely suffers from omitted variable problems. One possible solution is to try to control for more factors, such as age distribution, gender distribution, education levels, law enforcement efforts, and so on, in a multiple regression analysis. But many factors might be hard to control for. In Chapter 9, we showed how including the *crmrte* from a previous year—in this case, 1982—can help to control for the fact that different cities have historically different crime rates. This is one way to use two years of data for estimating a causal effect.

An alternative way to use panel data is to view the unobserved factors affecting the dependent variable as consisting of two types: those that are constant and those that vary over time. Letting i denote the cross-sectional unit and t the time period, we can write a model with a single observed explanatory variable as

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it},\ t = 1, 2. \qquad [13.20]$$

In the notation y_{it}, i denotes the person, firm, city, and so on, and t denotes the time period. The variable $d2_t$ is a dummy variable that equals zero when $t = 1$ and one when $t = 2$; it does not change across i, which is why it has no i subscript. Therefore, the intercept for $t = 1$ is β_0, and the intercept for $t = 2$ is $\beta_0 + \delta_0$. Just as in using independently pooled cross sections, allowing the intercept to change over time is important in most applications. In the crime example, secular trends in the United States will cause crime rates in all U.S. cities to change, perhaps markedly, over a five-year period.

The variable a_i captures all unobserved, time-constant factors that affect y_{it}. (The fact that a_i has no t subscript tells us that it does not change over time.) Generically, a_i is called an **unobserved effect**. It is also common in applied work to find a_i referred to as a **fixed effect**, which helps us to

remember that a_i is fixed over time. The model in (13.20) is called an **unobserved effects model** or a **fixed effects model**. In applications, you might see a_i referred to as **unobserved heterogeneity** as well (or *individual heterogeneity*, *firm heterogeneity*, *city heterogeneity*, and so on).

The error u_{it} is often called the **idiosyncratic error** or time-varying error, because it represents unobserved factors that change over time and affect y_{it}. These are very much like the errors in a straight time series regression equation.

A simple unobserved effects model for city crime rates for 1982 and 1987 is

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + u_{it}, \qquad [13.21]$$

where *d87* is a dummy variable for 1987. Because *i* denotes different cities, we call a_i an *unobserved city effect* or a *city fixed effect*: it represents all factors affecting city crime rates that do not change over time. Geographical features, such as the city's location in the United States, are included in a_i. Many other factors may not be exactly constant, but they might be roughly constant over a five-year period. These might include certain demographic features of the population (age, race, and education). Different cities may have their own methods for reporting crimes, and the people living in the cities might have different attitudes toward crime; these are typically slow to change. For historical reasons, cities can have very different crime rates, and historical factors are effectively captured by the unobserved effect a_i.

How should we estimate the parameter of interest, β_1, given two years of panel data? One possibility is just to pool the two years and use OLS, essentially as in Section 13-1. This method has two drawbacks. The most important of these is that, in order for pooled OLS to produce a consistent estimator of β_1, we would have to assume that the unobserved effect, a_i, is uncorrelated with x_{it}. We can easily see this by writing (13.20) as

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + v_{it}, \quad t = 1, 2, \qquad [13.22]$$

where $v_{it} = a_i + u_{it}$ is often called the **composite error**. From what we know about OLS, we must assume that v_{it} is uncorrelated with x_{it}, where $t = 1$ or 2, for OLS to estimate β_1 (and the other parameters consistently). This is true whether we use a single cross section or pool the two cross sections. Therefore, even if we assume that the idiosyncratic error u_{it} is uncorrelated with x_{it}, pooled OLS is biased and inconsistent if a_i and x_{it} are correlated. The resulting bias in pooled OLS is sometimes called **heterogeneity bias**, but it is really just bias caused from omitting a time-constant variable.

GOING FURTHER 13.3

Suppose that a_i, u_{i1}, and u_{i2} have zero means and are pairwise uncorrelated. Show that $\text{Cov}(v_{i1}, v_{i2}) = \text{Var}(a_i)$, so that the composite errors are positively serially correlated across time, unless $a_i = 0$. What does this imply about the usual OLS standard errors from pooled OLS estimation?

To illustrate what happens, we use the data in CRIME2 to estimate (13.21) by pooled OLS. Because there are 46 cities and two years for each city, there are 92 total observations:

$$\widehat{crmrte} = 93.42 + 7.94\, d87 + .427\, unem \qquad [13.23]$$
$$(12.74)\ (7.98)\quad (1.188)$$
$$n = 92,\ R^2 = .012.$$

(When reporting the estimated equation, we usually drop the *i* and *t* subscripts.) The coefficient on *unem*, though positive in (13.23), has a very small *t* statistic. Thus, using pooled OLS on the two years has not substantially changed anything from using a single cross section. This is not surprising because using pooled OLS does not solve the omitted variables problem. (The standard errors in this equation are incorrect because of the serial correlation described in Going Further 13.3, but we ignore this as pooled OLS is not the focus here.)

In most applications, the main reason for collecting panel data is to allow for the unobserved effect, a_i, to be correlated with the explanatory variables. For example, in the crime equation, we want to allow

the unmeasured city factors in a_i that affect the crime rate also to be correlated with the unemployment rate. It turns out that this is simple to allow: because a_i is constant over time, we can difference the data across the two years. More precisely, for a cross-sectional observation i, write the two years as

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \quad (t = 2)$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \quad (t = 1).$$

If we subtract the *second* equation from the *first*, we obtain

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1}),$$

or

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i, \qquad [13.24]$$

where Δ denotes the change from $t = 1$ to $t = 2$. The unobserved effect, a_i, does not appear in (13.24): it has been "differenced away." Also, the intercept in (13.24) is actually the *change* in the intercept from $t = 1$ to $t = 2$ as given in equation (13.22).

Equation (13.24), which we call the **first-differenced equation**, is very simple. It is just a single cross-sectional equation, but each variable is differenced over time. We can analyze (13.24) using the methods we developed in Part 1, provided the key assumptions are satisfied. The most important of these is that Δu_i is uncorrelated with Δx_i. This assumption holds if the idiosyncratic error at each time t, u_{it}, is uncorrelated with the explanatory variable in *both* time periods. This is another version of the **strict exogeneity** assumption that we encountered in Chapter 10 for time series models. In particular, this assumption rules out the case in which x_{it} is the lagged dependent variable, $y_{i,t-1}$. Unlike in Chapter 10, we allow x_{it} to be correlated with unobservables that are constant over time. When we obtain the OLS estimator of β_1 from (13.24), we call the resulting estimator the **first-differenced estimator**.

In the crime example, assuming that Δu_i and $\Delta unem_i$ are uncorrelated may be reasonable, but it can also fail. For example, suppose that law enforcement effort (which is in the idiosyncratic error) increases more in cities where the unemployment rate decreases. This can cause negative correlation between Δu_i and $\Delta unem_i$, which would then lead to bias in the OLS estimator. Naturally, this problem can be overcome to some extent by including more factors in the equation, something we will cover later. As usual, it is always possible that we have not accounted for enough time-varying factors.

Another crucial condition is that Δx_i must have some variation across i. This qualification fails if the explanatory variable does not change over time for any cross-sectional observation, or if it changes by the same amount for every observation. This is not an issue in the crime rate example because the unemployment rate changes across time for almost all cities. But, if i denotes an individual and x_{it} is a dummy variable for gender, $\Delta x_i = 0$ for all i; we clearly cannot estimate (13.24) by OLS in this case. This actually makes perfectly good sense: because we allow a_i to be correlated with x_{it}, we cannot hope to separate the effect of a_i on y_{it} from the effect of any variable that does not change over time.

The only other assumption we need to apply to the usual OLS statistics is that (13.24) satisfies the homoskedasticity assumption. This is reasonable in many cases, and, if it does not hold, we know how to test and correct for heteroskedasticity using the methods in Chapter 8. It is sometimes fair to assume that (13.24) fulfills all of the classical linear model assumptions. The OLS estimators are unbiased and all statistical inference is exact in such cases.

When we estimate (13.24) for the crime rate example, we get

$$\widehat{\Delta crmrte} = 15.40 + 2.22\ \Delta unem$$
$$(4.70) \quad (.88) \qquad [13.25]$$
$$n = 46,\ R^2 = .127,$$

which now gives a positive, statistically significant relationship between the crime and unemployment rates. Thus, differencing to eliminate time-constant effects makes a big difference in this example. The

intercept in (13.25) also reveals something interesting. Even if $\Delta unem = 0$, we predict an increase in the crime rate (crimes per 1,000 people) of 15.40. This reflects a secular increase in crime rates throughout the United States from 1982 to 1987.

Even if we do not begin with the unobserved effects model (13.20), using differences across time makes intuitive sense. Rather than estimating a standard cross-sectional relationship—which may suffer from omitted variables, thereby making ceteris paribus conclusions difficult—equation (13.24) explicitly considers how changes in the explanatory variable over time affect the change in y over the same time period. Nevertheless, it is still very useful to have (13.20) in mind: it explicitly shows that we can estimate the effect of x_{it} on y_{it}, holding a_i fixed.

Although differencing two years of panel data is a powerful way to control for unobserved effects, it is not without cost. First, panel data sets are harder to collect than a single cross section, especially for individuals. We must use a survey and keep track of the individual for a follow-up survey. It is often difficult to locate some people for a second survey. For units such as firms, some will go bankrupt or merge with other firms. Panel data are much easier to obtain for schools, cities, counties, states, and countries.

Even if we have collected a panel data set, the differencing used to eliminate a_i can greatly reduce the variation in the explanatory variables. While x_{it} frequently has substantial variation in the cross section for each t, Δx_i may not have much variation. We know from Chapter 3 that a little variation in Δx_i can lead to a large standard error for $\hat{\beta}_1$ when estimating (13.24) by OLS. We can combat this by using a large cross section, but this is not always possible. Also, using longer differences over time is sometimes better than using year-to-year changes.

As an example, consider the problem of estimating the return to education, now using panel data on individuals for two years. The model for person i is

$$\log(wage_{it}) = \beta_0 + \delta_0 d2_t + \beta_1 educ_{it} + a_i + u_{it}, \quad t = 1, 2,$$

where a_i contains unobserved ability—which is probably correlated with $educ_{it}$. Again, we allow different intercepts across time to account for aggregate productivity gains (and inflation, if $wage_{it}$ is in nominal terms). Because, by definition, innate ability does not change over time, panel data methods seem ideally suited to estimate the return to education. The equation in first differences is

$$\Delta\log(wage_i) = \delta_0 + \beta_1 \Delta educ_i + \Delta u_i, \qquad [13.26]$$

and we can estimate this by OLS. The problem is that we are interested in working adults, and for most employed individuals, education does not change over time. If only a small fraction of our sample has $\Delta educ_i$ different from zero, it will be difficult to get a precise estimator of β_1 from (13.26), unless we have a rather large sample size. In theory, using a first-differenced equation to estimate the return to education is a good idea, but it does not work very well with most currently available panel data sets.

Adding several explanatory variables causes no difficulties. We begin with the unobserved effects model

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \qquad [13.27]$$

for $t = 1$ and 2. This equation looks more complicated than it is because each explanatory variable has three subscripts. The first denotes the cross-sectional observation number, the second denotes the time period, and the third is just a variable label.

EXAMPLE 13.5 Sleeping versus Working

We use the two years of panel data in SLP75_81, from Biddle and Hamermesh (1990), to estimate the tradeoff between sleeping and working. In Problem 3 in Chapter 3, we used just the 1975 cross section. The panel data set for 1975 and 1981 has 239 people, which is much smaller than the 1975 cross

section that includes over 700 people. An unobserved effects model for total minutes of sleeping per week is

$$slpnap_{it} = \beta_0 + \delta_0 d81_t + \beta_1 totwrk_{it} + \beta_2 educ_{it} + \beta_3 marr_{it} + \beta_4 yngkid_{it} + \beta_5 gdhlth_{it} + a_i + u_{it}, \quad t = 1, 2.$$

The unobserved effect, a_i, would be called an *unobserved individual effect* or an *individual fixed effect*. It is potentially important to allow a_i to be correlated with $totwrk_{it}$: the same factors (some biological) that cause people to sleep more or less (captured in a_i) are likely correlated with the amount of time spent working. Some people just have more energy, and this causes them to sleep less and work more. The variable *educ* is years of education, *marr* is a marriage dummy variable, *yngkid* is a dummy variable indicating the presence of a small child, and *gdhlth* is a "good health" dummy variable. Notice that we do not include gender or race (as we did in the cross-sectional analysis), because these do not change over time; they are part of a_i. Our primary interest is in β_1.

Differencing across the two years gives the estimable equation

$$\Delta slpnap_i = \delta_0 + \beta_1 \Delta totwrk_i + \beta_2 \Delta educ_i + \beta_3 \Delta marr_i + \beta_4 \Delta yngkid_i + \beta_5 \Delta gdhlth_i + \Delta u_i.$$

Assuming that the change in the idiosyncratic error, Δu_i, is uncorrelated with the changes in all explanatory variables, we can get consistent estimators using OLS. This gives

$$\widehat{\Delta slpnap} = \underset{(45.87)}{-92.63} - \underset{(.036)}{.227}\,\Delta totwrk - \underset{(48.759)}{.024}\,\Delta educ + \underset{(92.86)}{104.21}\,\Delta marr + \underset{(87.65)}{94.67}\,\Delta yngkid + \underset{(76.60)}{87.58}\,\Delta gdhlth \qquad [13.28]$$

$$n = 239, R^2 = .150.$$

The coefficient on $\Delta totwrk$ indicates a tradeoff between sleeping and working: holding other factors fixed, one more hour of work is associated with $.227(60) = 13.62$ fewer minutes of sleeping. The t statistic (-6.31) is very significant. No other estimates, except the intercept, are statistically different from zero. The F test for joint significance of all variables except $\Delta totwrk$ gives p-value = .49, which means they are jointly insignificant at any reasonable significance level and could be dropped from the equation.

The standard error on $\Delta educ$ is especially large relative to the estimate. This is the phenomenon described earlier for the wage equation. In the sample of 239 people, 183 (76.6%) have no change in education over the six-year period; 90% of the people have a change in education of at most one year. As reflected by the extremely large standard error of $\hat{\beta}_2$, there is not nearly enough variation in education to estimate β_2 with any precision. Anyway, $\hat{\beta}_2$ is practically very small.

Panel data can also be used to estimate finite distributed lag models. Even if we specify the equation for only two years, we need to collect more years of data to obtain the lagged explanatory variables. The following is a simple example.

EXAMPLE 13.6 Distributed Lag of Crime Rate on Clear-Up Rate

Eide (1994) uses panel data from police districts in Norway to estimate a distributed lag model for crime rates. The single explanatory variable is the "clear-up percentage" (*clrprc*)—the percentage of crimes that led to a conviction. The crime rate data are from the years 1972 and 1978. Following

Eide, we lag *clrprc* for one and two years: it is likely that past clear-up rates have a deterrent effect on current crime. This leads to the following unobserved effects model for the two years:

$$\log(crime_{it}) = \beta_0 + \delta_0 d78_t + \beta_1 clrprc_{i,t-1} + \beta_2 clrprc_{i,t-2} + a_t + u_{it}.$$

When we difference the equation and estimate it using the data in CRIME3, we get

$$\widehat{\Delta \log(crime)} = .086 - .0040\,\Delta clrprc_{-1} - .0132\,\Delta clrprc_{-2} \quad [13.29]$$
$$(.064)\ (.0047)\qquad (.0052)$$
$$n = 53,\ R^2 = .193,\ \bar{R}^2 = .161.$$

The second lag is negative and statistically significant, which implies that a higher clear-up percentage two years ago would deter crime this year. In particular, a 10 percentage point increase in *clrprc* two years ago would lead to an estimated 13.2% drop in the crime rate this year. This suggests that using more resources for solving crimes and obtaining convictions can reduce crime in the future.

13-3a Organizing Panel Data

In using panel data in an econometric study, it is important to know how the data should be stored. We must be careful to arrange the data so that the different time periods for the same cross-sectional unit (person, firm, city, and so on) are easily linked. For concreteness, suppose that the data set is on cities for two different years. For most purposes, the best way to enter the data is to have *two* records for each city, one for each year: the first record for each city corresponds to the early year, and the second record is for the later year. These two records should be adjacent. Therefore, a data set for 100 cities and two years will contain 200 records. The first two records are for the first city in the sample, the next two records are for the second city, and so on. (See Table 1.5 in Chapter 1 for an example.) This makes it easy to construct the differences to store these in the second record for each city and to do a pooled cross-sectional analysis, which can be compared with the differencing estimation.

Most of the two-period panel data sets accompanying this text are stored in this way (for example, CRIME2, CRIME3, GPA3, LOWBRTH, and RENTAL). We use a direct extension of this scheme for panel data sets with more than two time periods.

A second way of organizing two periods of panel data is to have only one record per cross-sectional unit. This requires two entries for each variable, one for each time period. The panel data in SLP75_81 are organized in this way. Each individual has data on the variables *slpnap75*, *slpnap81*, *totwrk75*, *totwrk81*, and so on. Creating the differences from 1975 to 1981 is easy. Other panel data sets with this structure are TRAFFIC1 and VOTE2. Putting the data in one record, however, does not allow a pooled OLS analysis using the two time periods on the original data. Also, this organizational method does not work for panel data sets with more than two time periods, a case we will consider in Section 13-5.

13-4 Policy Analysis with Two-Period Panel Data

Panel data sets are very useful for policy analysis and, in particular, program evaluation. In the simplest program evaluation setup, a sample of individuals, firms, cities, and so on, is obtained in the first time period. Some of these units, those in the treatment group, then take part in a particular program in a later time period; the ones that do not are the control group. This is similar to the natural experiment literature discussed earlier, with one important difference: the *same* cross-sectional units appear in each time period.

As an example, suppose we wish to evaluate the effect of a Michigan job training program on worker productivity of manufacturing firms (see also Computer Exercise C3 in Chapter 9). Let $scrap_{it}$ denote the scrap rate of firm i during year t (the number of items, per 100, that must be scrapped due

to defects). Let $grant_{it}$ be a binary indicator equal to one if firm i in year t received a job training grant. For the years 1987 and 1988, the model is

$$scrap_{it} = \beta_0 + \delta_0 y88_t + \beta_1 grant_{it} + a_i + u_{it}, \quad t = 1, 2, \tag{13.30}$$

where $y88_t$ is a dummy variable for 1988 and a_i is the *unobserved firm effect* or the *firm fixed effect*. The unobserved effect contains such factors as average employee ability, capital, and managerial skill; these are roughly constant over a two-year period. We are concerned about a_i being systematically related to whether a firm receives a grant. For example, administrators of the program might give priority to firms whose workers have lower skills. Or, the opposite problem could occur: to make the job training program appear effective, administrators may give the grants to employers with more productive workers. Actually, in this particular program, grants were awarded on a first-come, first-served basis. But whether a firm applied early for a grant could be correlated with worker productivity. In that case, an analysis using a single cross section or just a pooling of the cross sections will produce biased and inconsistent estimators.

Differencing to remove a_i gives

$$\Delta scrap_i = \delta_0 + \beta_1 \Delta grant_i + \Delta u_i. \tag{13.31}$$

Therefore, we simply regress the change in the scrap rate on the change in the grant indicator. Because no firms received grants in 1987, $grant_{i1} = 0$ for all i, and so $\Delta grant_i = grant_{i2} - grant_{i1} = grant_{i2}$, which simply indicates whether the firm received a grant in 1988. However, it is generally important to difference all variables (dummy variables included) because this is necessary for removing a_i in the unobserved effects model (13.30).

Estimating the first-differenced equation using the data in JTRAIN gives

$$\widehat{\Delta scrap} = -.564 - .739\ \Delta grant$$
$$(.405)\quad (.683)$$
$$n = 54,\ R^2 = .022.$$

Therefore, we estimate that having a job training grant lowered the scrap rate on average by $-.739$. But the estimate is not statistically different from zero.

We get stronger results by using $\log(scrap)$ and estimating the percentage effect:

$$\widehat{\Delta \log(scrap)} = -.057 - .317 \Delta grant$$
$$(.097)\quad (.164)$$
$$n = 54,\ R^2 = .067.$$

Having a job training grant is estimated to lower the scrap rate by about 27.2%. [We obtain this estimate from equation (7.10): $\exp(-.317) - 1 \approx -.272$.] The t statistic is about -1.93, which is marginally significant. By contrast, using pooled OLS of $\log(scrap)$ on $y88$ and $grant$ gives $\hat{\beta}_1 = .057$ (standard error $= .431$). Thus, we find no significant relationship between the scrap rate and the job training grant. Because this differs so much from the first-difference estimates, it suggests that firms that have lower-ability workers are more likely to receive a grant.

It is useful to study the program evaluation model more generally. Let y_{it} denote an outcome variable and let $prog_{it}$ be a program participation dummy variable. The simplest unobserved effects model is

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + a_i + u_{it}. \tag{13.32}$$

If program participation only occurred in the second period, then the OLS estimator of β_1 in the differenced equation has a very simple representation:

$$\hat{\beta}_1 = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}. \tag{13.33}$$

That is, we compute the average change in y over the two time periods for the treatment and control groups. Then, $\hat{\beta}_1$ is the difference of these. This is the panel data version of the difference-in-differences estimator in equation (13.11) for two pooled cross sections. With panel data, we have a potentially important advantage: we can difference y across time for the *same* cross-sectional units. This allows us to control for person-, firm-, or city-specific effects, as the model in (13.32) makes clear.

If program participation takes place in both periods, $\hat{\beta}_1$ cannot be written as in (13.33), but we interpret it in the same way: it is the change in the average value of y due to program participation.

Controlling for time-varying factors does not change anything of significance. We simply difference those variables and include them along with $\Delta prog$. This allows us to control for time-varying variables that might be correlated with program designation.

The same differencing method works for analyzing the effects of any policy that varies across city or state. The following is a simple example.

EXAMPLE 13.7 Effect of Drunk Driving Laws on Traffic Fatalities

Many states in the United States have adopted different policies in an attempt to curb drunk driving. Two types of laws that we will study here are *open container laws*—which make it illegal for passengers to have open containers of alcoholic beverages, and *administrative per se laws*—which allow courts to suspend licenses after a driver is arrested for drunk driving but before the driver is convicted. One possible analysis is to use a single cross section of states to regress driving fatalities (or those related to drunk driving) on dummy variable indicators for whether each law is present. This is unlikely to work well because states decide, through legislative processes, whether they need such laws. Therefore, the presence of laws is likely to be related to the average drunk driving fatalities in recent years. A more convincing analysis uses panel data over a time period during which some states adopted new laws (and some states may have repealed existing laws). The file TRAFFIC1 contains data for 1985 and 1990 for all 50 states and the District of Columbia. The dependent variable is the number of traffic deaths per 100 million miles driven (*dthrte*). In 1985, 19 states had open container laws, while 22 states had such laws in 1990. In 1985, 21 states had per se laws; the number had grown to 29 by 1990.

Using OLS after first differencing gives

$$\widehat{\Delta dthrte} = -.497 - .420\,\Delta open - .151\Delta admn \quad [13.34]$$
$$(.052)\ (.206)\qquad (.117)$$
$$n = 51,\ R^2 = .119.$$

GOING FURTHER 13.4

In Example 13.7, $\Delta admn = -1$ for the state of Washington. Explain what this means.

The estimates suggest that adopting an open container law lowered the traffic fatality rate by .42, a nontrivial effect given that the average death rate in 1985 was 2.7 with a standard deviation of about .6. The estimate is statistically significant at the 5% level against a two-sided alternative. The administrative per se law has a smaller effect, and its t statistic is only -1.29; but the estimate is the sign we expect. The intercept in this equation shows that traffic fatalities fell substantially for all states over the five-year period, whether or not there were any law changes. The states that adopted an open container law over this period saw a further drop, on average, in fatality rates.

Other laws might also affect traffic fatalities, such as seat belt laws, motorcycle helmet laws, and maximum speed limits. In addition, we might want to control for age and gender distributions, as well as measures of how influential an organization such as Mothers Against Drunk Driving is in each state.

13-5 Differencing with More Than Two Time Periods

We can also use differencing with more than two time periods. For illustration, suppose we have N individuals and $T = 3$ time periods for each individual. A general unobserved effects model is

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \quad [13.35]$$

for $t = 1, 2$, and 3. (The total number of observations is therefore $3N$.) Notice that we now include two time-period dummies in addition to the intercept. It is a good idea to allow a separate intercept for each time period, especially when we have a small number of them. The base period, as always, is $t = 1$. The intercept for the second time period is $\delta_1 + \delta_2$, and so on. We are primarily interested in $\beta_1, \beta_2, \ldots, \beta_k$. If the unobserved effect a_i is correlated with any of the explanatory variables, then using pooled OLS on the three years of data results in biased and inconsistent estimates.

The key assumption is that the idiosyncratic errors are uncorrelated with the explanatory variable in each time period:

$$\text{Cov}(x_{itj}, u_{is}) = 0, \quad \text{for all, } t, s, \text{ and } j. \quad [13.36]$$

That is, the explanatory variables are *strictly exogenous* after we take out the unobserved effect, a_i. (The strict exogeneity assumption stated in terms of a zero conditional expectation is given in the chapter appendix.) Assumption (13.36) rules out cases in which future explanatory variables react to current changes in the idiosyncratic errors, as must be the case if x_{itj} is a lagged dependent variable. If we have omitted an important time-varying variable, then (13.36) is generally violated. Measurement error in one or more explanatory variables can cause (13.36) to be false, just as in Chapter 9. In Chapters 15 and 16, we will discuss what can be done in such cases.

If a_i is correlated with x_{itj}, then x_{itj} will be correlated with the *composite* error, $v_{it} = a_i + u_{it}$, under (13.36). We can eliminate a_i by differencing adjacent periods. In the $T = 3$ case, we subtract time period one from time period two and time period two from time period three. This gives

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \Delta u_{it}, \quad [13.37]$$

for $t = 2$ and 3. We do not have a differenced equation for $t = 1$ because there is nothing to subtract from the $t = 1$ equation. Now, (13.37) represents *two* time periods for each individual in the sample. If this equation satisfies the classical linear model assumptions, then pooled OLS gives unbiased estimators, and the usual t and F statistics are valid for hypothesis. We can also appeal to asymptotic results. The important requirement for OLS to be consistent is that Δu_{it} is uncorrelated with Δx_{itj} for all j and $t = 2$ and 3. This is the natural extension from the two time period case.

Notice how (13.37) contains the differences in the year dummies, $d2_t$ and $d3_t$. For $t = 2$, $\Delta d2_t = 1$ and $\Delta d3_t = 0$; for $t = 3$, $\Delta d2_t = -1$ and $\Delta d3_t = 1$. The intercept in (13.37) has been differenced away. This is inconvenient for the purpose of computing an R-squared, but there are two simple remedies. First, some regression packages allow you to compute the total sum of squares (SST) as if there is a constant, and this provides a better goodness-of-feet measure for explaining Δy_{it}. Second, one can estimate a simple transformation of the equation that includes an intercept:

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \Delta u_{it}, \; t = 2, 3.$$

The estimates of the β_j will not change, and now the R-squared is properly computed. One reason for estimating the equation in (13.37) is that we can compare the estimates directly with using pooled OLS on the levels and with methods that we cover in Chapter 14.

With more than three time periods, things are similar. If we have the same T time periods for each of N cross-sectional units, we say that the data set is a **balanced panel**: we have the same time periods for all individuals, firms, cities, and so on. When T is small relative to N, we should include a dummy variable for each time period to account for secular changes that are not being modeled. Therefore, after first differencing, the equation looks like

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \cdots + \delta_T \Delta dT_t + \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \Delta u_{it}, \; t = 2, 3, \ldots, T. \quad [13.38]$$

where we have $T - 1$ time periods on each unit i for the first-differenced equation. The total number of observations is $N(T - 1)$. To obtain a useful R-squared measure, we can instead estimate an intercept and include the time dummies $d_{3t}, d_{4t}, \ldots, dT_t$ in place of the differences in the dummies.

It is simple to estimate (13.38) by pooled OLS, provided the observations have been properly organized and the differencing carefully done. To facilitate first differencing, the data file should consist of NT records. The first T records are for the first cross-sectional observation, arranged chronologically; the second T records are for the second cross-sectional observations, arranged chronologically; and so on. Then, we compute the differences, with the change from $t - 1$ to t stored in the time t record. Therefore, the differences for $t = 1$ should be missing values for all N cross-sectional observations. Without doing this, you run the risk of using bogus observations in the regression analysis. An invalid observation is created when the last observation for, say, person $i - 1$ is subtracted from the first observation for person i. If you do the regression on the differenced data, and NT or $NT - 1$ observations are reported, then you forgot to set the $t = 1$ observations as missing.

When using more than two time periods, we must assume that Δu_{it} is uncorrelated over time for the usual standard errors and test statistics to be valid. This assumption is sometimes reasonable, but it does not follow if we assume that the original idiosyncratic errors, u_{it}, are uncorrelated over time (an assumption we will use in Chapter 14). In fact, if we assume the u_{it} are serially uncorrelated with constant variance, then the correlation between Δu_{it} and $\Delta u_{i,t+1}$ can be shown to be $-.5$. If u_{it} follows a stable AR(1) model, then Δu_{it} will be serially correlated. Only when u_{it} follows a random walk will Δu_{it} be serially uncorrelated.

It is easy to test for serial correlation in the first-differenced equation. Let $r_{it} = \Delta u_{it}$ denote the first difference of the original error. If r_{it} follows the AR(1) model $r_{it} = \rho r_{i,t-1} + e_{it}$, then we can easily test H_0: $\rho = 0$. First, we estimate (13.38) by pooled OLS and obtain the residuals, $\hat{r}_{it}$.

Then, we run a simple pooled OLS regression of $\hat{r}_{it}$ on $\hat{r}_{i,t-1}$, $t = 3, \ldots, T$, $i = 1, \ldots, N$, and compute a standard t test for the coefficient on $\hat{r}_{i,t-1}$. (Or we can make the t statistic robust to heteroskedasticity.) The coefficient $\hat{\rho}$ on $\hat{r}_{i,t-1}$ is a consistent estimator of ρ. Because we are using the lagged residual, we lose another time period. For example, if we started with $T = 3$, the differenced equation has two time periods, and the test for serial correlation is just a cross-sectional regression of the residuals from the third time period on the residuals from the second time period. We will give an example later.

If we detect serial correlation—and even if we do not bother to test for serial correlation—it is possible to adjust the standard errors to allow unrestricted forms of serial correlation and heteroskedasticity. Such methods, which fall under the topic of **cluster-robust standard errors**, are described in nontechnical terms in the appendix to this chapter, and a formal treatment is in Wooldridge (2010, Chapter 10). The standard approach assumes the observations are independent across i but, with a moderately large N size and not "too large" T, they allow any kind of serial correlation pattern (along with heteroskedasticity).

GOING FURTHER 13.5

Does serial correlation in Δu_{it} cause the first-differenced estimator to be biased and inconsistent? Why is serial correlation a concern?

Less common these days, but still useful, is to correct for the presence of AR(1) serial correlation in r_{it} by using feasible GLS. Essentially, within each cross-sectional observation, we would use the Prais-Winsten transformation based on $\hat{r}_{it}$ described in the previous paragraph. (We clearly prefer Prais-Winsten to Cochrane-Orcutt here, as dropping the first time period would now mean losing N cross-sectional observations.) Unfortunately, commands that perform AR(1) corrections for time series regressions generally will not work when applied to time series data. Standard Prais-Winsten methods will treat the observations as if they followed an AR(1) process across i and t; this makes no sense, as we are assuming the observations are independent across i. Wooldridge (2010, Chapter 10) discusses how one can use GLS methods in first differenced equations, and some software packages have special commands that perform the estimation.

If there is no serial correlation in the errors, the usual methods for dealing with heteroskedasticity are valid. We can use the usual heteroskedasticity-robust standard errors as the simplest fix. Also, we can use the Breusch-Pagan and White tests for heteroskedasticity from Chapter 8.

Differencing more than two years of panel data is very useful for policy analysis, as shown by the following example.

EXAMPLE 13.8 **Effect of Enterprise Zones on Unemployment Claims**

Papke (1994) studied the effect of the Indiana enterprise zone (EZ) program on unemployment claims. She analyzed 22 cities in Indiana over the period from 1980 to 1988. Six enterprise zones were designated in 1984, and four more were assigned in 1985. Twelve of the cities in the sample did not receive an enterprise zone over this period; they served as the control group.

A simple policy evaluation model is

$$\log(uclms_{it}) = \theta_t + \beta_1 ez_{it} + a_i + u_{it},$$

where $uclms_{it}$ is the number of unemployment claims filed during year t in city i. The parameter θ_t just denotes a different intercept for each time period. Generally, unemployment claims were falling statewide over this period, and this should be reflected in the different year intercepts. The binary variable ez_{it} is equal to one if city i at time t was an enterprise zone; we are interested in β_1. The unobserved effect a_i represents fixed factors that affect the economic climate in city i. Because enterprise zone designation was not determined randomly—enterprise zones are usually economically depressed areas—it is likely that ez_{it} and a_i are positively correlated (high a_i means higher unemployment claims, which lead to a higher chance of being given an EZ). Thus, we should difference the equation to eliminate a_i:

$$\Delta\log(uclms_{it}) = \delta_1 \Delta d81_t + \delta_2 \Delta d82_t + \cdots + \delta_8 \Delta d88_t + \beta_1 \Delta ez_{it} + \Delta u_{it}. \qquad [13.39]$$

The dependent variable in this equation, the change in $\log(uclms_{it})$, is the approximate annual growth rate in unemployment claims from year $t-1$ to t. We can estimate this equation for the years 1981 to 1988 using the data in EZUNEM; the total sample size is $22 \cdot 8 = 176$. The estimate of β_1 is $\hat{\beta}_1 = -.182$ (standard error $= .078$). Therefore, it appears that the presence of an EZ causes about a 16.6% $[\exp(-.182) - 1 \approx -.166]$ fall in unemployment claims. This is an economically large and statistically significant effect.

If we add the lagged OLS residuals to the differenced equation (and lose the year 1981), we get $\hat{\rho} = -.197$ $(t = -2.44)$, so there is evidence of some negative serial correlation. When we compute a standard error on the *ez* dummy variable that is robust to both serial correlation, as described in the appendix, it is .092, which is above the usual OLS standard error reported above. The cluster-robust t statistic is about −1. 98, and so the estimated enterprise zone is less statistically significant.

EXAMPLE 13.9 **County Crime Rates in North Carolina**

Cornwell and Trumbull (1994) used data on 90 counties in North Carolina, for the years 1981 through 1987, to estimate an unobserved effects model of crime; the data are contained in CRIME4. Here, we estimate a simpler version of their model, and we difference the equation over time to eliminate a_i, the unobserved effect. (Cornwell and Trumbull use a different transformation, which we will cover in Chapter 14.) Various factors including geographical location, attitudes toward crime, historical records, and reporting conventions might be contained in a_i. The crime rate is number of crimes per person, *prbarr* is the estimated probability of arrest, *prbconv* is the estimated probability of conviction (given an arrest), *prbpris* is the probability of serving time in prison (given a conviction), *avgsen* is the average sentence length served, and *polpc* is the number of police officers per capita. As is standard in criminometric studies, we use the logs of all variables to estimate elasticities. We also include a full set of year dummies to control for state trends in crime rates. We can use the years 1982 through 1987 to estimate the differenced equation. The quantities in parentheses are the usual OLS

standard errors; the quantities in brackets are standard errors robust to both serial correlation and heteroskedasticity:

$$\begin{aligned}\widehat{\Delta\log(crmrte)} &= .008 - .100\,d83 - .048\,d84 - .005\,d85\\ &\quad (.017)\ (.024)\quad (.024)\quad (.023)\\ &\quad [.014]\ [.022]\quad [.020]\quad [.025]\\ &\quad + .028\,d86 + .041\,d87 - .327\,\Delta\log(prbarr)\\ &\quad (.024)\quad (.024)\quad (.030)\\ &\quad [.021]\quad [.024]\quad [.056]\\ &\quad - .238\,\Delta\log(prbconv) - .165\,\Delta\log(prbpris)\\ &\quad (.018)\quad (.026)\\ &\quad [.040]\quad [.046]\\ &\quad - .022\,\Delta\log(avgsen) + .398\,\Delta\log(polpc)\\ &\quad (.022)\quad (.027)\\ &\quad [.026]\quad [.103]\\ n &= 540,\ R^2 = .433,\ \overline{R}^2 = .422.\end{aligned} \qquad [13.40]$$

The three probability variables—of arrest, conviction, and serving prison time—all have the expected sign, and all are statistically significant. For example, a 1% increase in the probability of arrest is predicted to lower the crime rate by about .33%. The average sentence variable shows a modest deterrent effect, but it is not statistically significant.

The coefficient on the police per capita variable is somewhat surprising and is a feature of most studies that seek to explain crime rates. Interpreted causally, it says that a 1% increase in police per capita *increases* crime rates by about .4%. (The usual t statistic is very large, almost 15.) It is hard to believe that having more police officers causes more crime. What is going on here? There are at least two possibilities. First, the crime rate variable is calculated from *reported* crimes. It might be that, when there are additional police, more crimes are reported. Second, the police variable might be endogenous in the equation for other reasons: counties may enlarge the police force when they expect crime rates to increase. In this case, (13.33) cannot be interpreted in a causal fashion. In Chapters 15 and 16, we will cover models and estimation methods that can account for this additional form of endogeneity.

The special case of the White test for heteroskedasticity in Section 8-3 gives $F = 75.48$ and p-value $= .0000$, so there is strong evidence of heteroskedasticity. (Technically, this test is not valid if there is also serial correlation, but it is strongly suggestive.) Testing for AR(1) serial correlation yields $\hat{\rho} = -.233$, $t = -4.77$, so negative serial correlation exists. The standard errors in brackets adjust for serial correlation and heteroskedasticity. [See the discussion in the appendix.] No variables lose statistical significance, but the t statistics on the significant deterrent variables get notably smaller. For example, the t statistic on the probability of conviction variable goes from -13.22 using the usual OLS standard error to -6.10 using the fully robust standard error. Equivalently, the confidence intervals constructed using the robust standard errors will, appropriately, be much wider than those based on the usual OLS standard errors.

Naturally, we can apply the Chow test to panel data models estimated by first differencing. As in the case of pooled cross sections, we rarely want to test whether the intercepts are constant over time; for many reasons, we expect the intercepts to be different. Much more interesting is to test whether slope coefficients have changed over time, and we can easily carry out such tests by interacting the explanatory variables of interest with time-period dummy variables. Interestingly, while we cannot estimate the slopes on variables that do not change over time, we can test whether the partial effects of time-constant variables have changed over time. As an illustration, suppose we observe three years of

data on a random sample of people working in 2000, 2002, and 2004, and specify the model (for the log of wage, *lwage*),

$$lwage_{it} = \beta_0 + \delta_1 d02_t + \delta_2 d04_t + \beta_1 female_i + \gamma_1 d02_t female_i + \gamma_2 d04_t female_i + \mathbf{z}_{it}\boldsymbol{\lambda} + a_i + u_{it},$$

where $\mathbf{z}_{it}\boldsymbol{\lambda}$ is shorthand for other explanatory variables included in the model and their coefficients. When we first difference, we eliminate the intercept for 2000, β_0, and also the gender wage gap for 2000, β_1. However, the change in $d01_t female_i$ is $(\Delta d01_t)female_i$, which does not drop out. Consequently, we can estimate how the wage gap has changed in 2002 and 2004 relative to 2000, and we can test whether $\gamma_1 = 0$, or $\gamma_2 = 0$, or both. We might also ask whether the union wage premium has changed over time, in which case we include in the model $union_{it}$, $d02_t union_{it}$, and $d04_t union_{it}$. The coefficients on all of these explanatory variables can be estimated because $union_{it}$ would presumably have some time variation.

If one tries to estimate a model containing interactions by differencing by hand, it can be a bit tricky. For example, in the previous equation with union status, we must simply difference the interaction terms, $d02_t union_{it}$ and $d04_t union_{it}$. We *cannot* compute the proper differences as, say, $d02_t \Delta union_{it}$ and $d04_t \Delta union_{it}$, or by even replacing $d02_t$ and $d04_t$ with their first differences.

As a general comment, it is important to return to the original model and remember that the differencing is used to eliminate a_i. It is easiest to use a built-in command that allows first differencing as an option in panel data analysis, as discussed in the appendix to this chapter. (We will see some of the other options in Chapter 14.)

13-5a Potential Pitfalls in First Differencing Panel Data

In this and previous sections, we have argued that differencing panel data over time, in order to eliminate a time-constant unobserved effect, is a valuable method for obtaining causal effects. Nevertheless, differencing is not free of difficulties. We have already discussed potential problems with the method when the key explanatory variables do not vary much over time (and the method is useless for explanatory variables that never vary over time). Unfortunately, even when we do have sufficient time variation in the x_{itj}, first-differenced (FD) estimation can be subject to serious biases. We have already mentioned that strict exogeneity of the regressors is a critical assumption. Unfortunately, as discussed in Wooldridge (2010, Section 11-1), having more time periods generally does not reduce the inconsistency in the FD estimator when the regressors are not strictly exogenous (say, if $y_{i,t-1}$ is included among the x_{itj}).

Another important drawback to the FD estimator is that it can be worse than pooled OLS if one or more of the explanatory variables is subject to measurement error, especially the classical errors-in-variables model discussed in Section 9-3. Differencing a poorly measured regressor reduces its variation relative to its correlation with the differenced error caused by classical measurement error, resulting in a potentially sizable bias. Solving such problems can be very difficult. See Section 15-8 and Wooldridge (2010, Chapter 11).

Summary

We have studied methods for analyzing independently pooled cross-sectional and panel data sets. Independent cross sections arise when different random samples are obtained in different time periods (usually years). OLS using pooled data is the leading method of estimation, and the usual inference procedures are available, including corrections for heteroskedasticity. (Serial correlation is not an issue because the samples are independent across time.) Because of the time series dimension, we often allow different time intercepts. We might also interact time dummies with certain key variables to see how they have

changed over time. This is especially important in the policy evaluation literature for natural experiments. The difference-in-differences methodology, and its extensions, has proven very useful for studying policy interventions.

Panel data sets are being used more and more in applied work, especially for policy analysis. These are data sets in which the same cross-sectional units are followed over time. Panel data sets are most useful when controlling for time-constant unobserved features—of people, firms, cities, and so on—which we think might be correlated with the explanatory variables in our model. One way to remove the unobserved effect is to difference the data in adjacent time periods. Then, a standard OLS analysis on the differences can be used. Using two periods of data results in a cross-sectional regression of the differenced data. The usual inference procedures are asymptotically valid under homoskedasticity; exact inference is available under normality.

For more than two time periods, we can use pooled OLS on the differenced data; we lose the first time period because of the differencing. In addition to homoskedasticity, we must assume that the *differenced* errors are serially uncorrelated in order to apply the usual t and F statistics. (The chapter appendix contains a careful listing of the assumptions.) Naturally, any variable that is constant over time drops out of the analysis. The appendix contains a discussion of how one computes standard errors that allow for unrestricted forms of serial correlation and heteroskedasticity.

Key Terms

Average Treatment Effect
Balanced Panel
Clustering
Cluster-Robust Standard Errors
Composite Error
Difference-in-Differences (DD or DID) Estimator
Difference-in-Difference-in-Differences (DDD) Estimator
First-Differenced Equation
First-Differenced Estimator
Fixed Effect
Fixed Effects Model
Group-Specific
Heterogeneity Bias
Idiosyncratic Error
Independently Pooled Cross Section
Longitudinal Data
Natural Experiment
Panel Data
Parallel Trends Assumption
Quasi-Experiment
Strict Exogeneity
Unobserved Effect
Unobserved Effects Model
Unobserved Heterogeneity
Year Dummy Variables

Problems

1 In Example 13.1, assume that the averages of all factors other than *educ* have remained constant over time and that the average level of education is 12.2 for the 1972 sample and 13.3 in the 1984 sample. Using the estimates in Table 13.1, find the estimated change in average fertility between 1972 and 1984. (Be sure to account for the intercept change and the change in average education.)

2 Using the data in KIELMC, the following equations were estimated using the years 1978 and 1981:

$$\widehat{\log(price)} = 11.49 - .547\, nearinc + .394\, y81{\cdot}nearinc$$
$$(.26)\ (.058)\qquad (.080)$$
$$n = 321,\ R^2 = .220$$

and

$$\widehat{\log(price)} = 11.18 + .563\, y81 - .403\, y81{\cdot}nearinc$$
$$(.27)\ (.044)\qquad (.067)$$
$$n = 321,\ R^2 = .337.$$

Compare the estimates on the interaction term *y81·nearinc* with those from equation (13.9). Why are the estimates so different?

3 Why can we not use first differences when we have independent cross sections in two years (as opposed to panel data)?

4 If we think that β_1 is positive in (13.14) and that Δu_i and $\Delta unem_i$ are negatively correlated, what is the bias in the OLS estimator of β_1 in the first-differenced equation? [*Hint:* Review equation (5.4).]

5 Suppose that we want to estimate the effect of several variables on annual saving and that we have a panel data set on individuals collected on January 31, 1990, and January 31, 1992. If we include a year dummy for 1992 and use first differencing, can we also include age in the original model? Explain.

6 In 1985, neither Florida nor Georgia had laws banning open alcohol containers in vehicle passenger compartments. By 1990, Florida had passed such a law, but Georgia had not.

(i) Suppose you can collect random samples of the driving-age population in both states, for 1985 and 1990. Let *arrest* be a binary variable equal to unity if a person was arrested for drunk driving during the year. Without controlling for any other factors, write down a linear probability model that allows you to test whether the open container law reduced the probability of being arrested for drunk driving. Which coefficient in your model measures the effect of the law?

(ii) Why might you want to control for other factors in the model? What might some of these factors be?

(iii) Now, suppose that you can only collect data for 1985 and 1990 at the county level for the two states. The dependent variable would be the fraction of licensed drivers arrested for drunk driving during the year. How does this data structure differ from the individual-level data described in part (i)? What econometric method would you use?

7 (i) Using the data in INJURY for Kentucky, we find the estimated equation when *afchnge* is dropped from (13.13) is

$$\widehat{\log(durat)} = 1.129 + .253\, highearn + .198\, afchnge{\cdot}highearn$$
$$(0.022)\ \ (.042)\qquad\quad (.052)$$
$$n = 5{,}626;\ R^2 = .021.$$

Is it surprising that the estimate on the interaction is fairly close to that in (13.13)? Explain.

(ii) When *afchnge* is included but *highearn* is dropped, the result is

$$\widehat{\log(durat)} = 1.233 - .100\, afchnge + .447\, afchnge{\cdot}highearn$$
$$(0.023)\ \ (.040)\qquad\quad (.050)$$
$$n = 5{,}626;\ R^2 = .016.$$

Why is the coefficient on the interaction term now so much larger than in (13.13)? [*Hint:* In equation (13.10), what is the assumption being made about the treatment and control groups if $\beta_1 = 0$?]

Computer Exercises

C1 Use the data in FERTIL1 for this exercise.

(i) In the equation estimated in Example 13.1, test whether living environment at age 16 has an effect on fertility. (The base group is large city.) Report the value of the F statistic and the p-value.

(ii) Test whether region of the country at age 16 (South is the base group) has an effect on fertility.

(iii) Let u be the error term in the population equation. Suppose you think that the variance of u changes over time (but not with *educ*, *age*, and so on). A model that captures this is

$$u^2 = \gamma_0 + \gamma_1 y74 + \gamma_2 y76 + \cdots + \gamma_6 y84 + v.$$

Using this model, test for heteroskedasticity in u. (*Hint:* Your F test should have 6 and 1,122 degrees of freedom.)

(iv) Add the interaction terms $y74{\cdot}educ$, $y76{\cdot}educ$, . . . , $y84{\cdot}educ$ to the model estimated in Table 13.1. Explain what these terms represent. Are they jointly significant?

C2 Use the data in CPS78_85 for this exercise.

(i) How do you interpret the coefficient on *y85* in equation (13.2)? Does it have an interesting interpretation? (Be careful here; you must account for the interaction terms *y85·educ* and *y85·female*.)

(ii) Holding other factors fixed, what is the estimated percent increase in nominal wage for a male with 12 years of education? Propose a regression to obtain a confidence interval for this estimate. [*Hint:* To get the confidence interval, replace *y85·educ* with *y85·*(*educ* − 12); refer to Example 6.3.]

(iii) Reestimate equation (13.2) but let all wages be measured in 1978 dollars. In particular, define the real wage as *rwage* = *wage* for 1978 and as *rwage* = *wage*/1.65 for 1985. Now, use log(*rwage*) in place of log(*wage*) in estimating (13.2). Which coefficients differ from those in equation (13.2)?

(iv) Explain why the *R*-squared from your regression in part (iii) is not the same as in equation (13.2). (*Hint:* The residuals, and therefore the sum of squared residuals, from the two regressions *are* identical.)

(v) Describe how union participation changed from 1978 to 1985.

(vi) Starting with equation (13.2), test whether the union wage differential changed over time. (This should be a simple *t* test.)

(vii) Do your findings in parts (v) and (vi) conflict? Explain.

C3 Use the data in KIELMC for this exercise.

(i) The variable *dist* is the distance from each home to the incinerator site, in feet. Consider the model

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 \log(dist) + \delta_1 y81 \cdot \log(dist) + u.$$

If building the incinerator reduces the value of homes closer to the site, what is the sign of δ_1? What does it mean if $\beta_1 > 0$?

(ii) Estimate the model from part (i) and report the results in the usual form. Interpret the coefficient on $y81 \cdot \log(dist)$. What do you conclude?

(iii) Add *age*, age^2, *rooms*, *baths*, log(*intst*), log(*land*), and log(*area*) to the equation. Now, what do you conclude about the effect of the incinerator on housing values?

(iv) Why is the coefficient on log(*dist*) positive and statistically significant in part (ii) but not in part (iii)? What does this say about the controls used in part (iii)?

C4 Use the data in INJURY for this exercise.

(i) Using the data for Kentucky, reestimate equation (13.13), adding as explanatory variables *male*, *married*, and a full set of industry and injury type dummy variables. How does the estimate on *afchnge·highearn* change when these other factors are controlled for? Is the estimate still statistically significant?

(ii) What do you make of the small *R*-squared from part (i)? Does this mean the equation is useless?

(iii) Estimate equation (13.13) using the data for Michigan. Compare the estimates on the interaction term for Michigan and Kentucky. Is the Michigan estimate statistically significant? What do you make of this?

C5 Use the data in RENTAL for this exercise. The data for the years 1980 and 1990 include rental prices and other variables for college towns. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it},$$

where *pop* is city population, *avginc* is average income, and *pctstu* is student population as a percentage of city population (during the school year).

(i) Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable? What do you get for $\hat{\beta}_{pctstu}$?

(ii) Are the standard errors you report in part (i) valid? Explain.

(iii) Now, difference the equation and estimate by OLS. Compare your estimate of β_{pctstu} with that from part (ii). Does the relative size of the student population appear to affect rental prices?

(iv) Obtain the heteroskedasticity-robust standard errors for the first-differenced equation in part (iii). Does this change your conclusions?

C6 Use CRIME3 for this exercise.

(i) In the model of Example 13.6, test the hypothesis H_0: $\beta_1 = \beta_2$. (*Hint:* Define $\theta_1 = \beta_1 - \beta_2$ and write β_1 in terms of θ_1 and β_2. Substitute this into the equation and then rearrange. Do a t test on θ_1.)

(ii) If $\beta_1 = \beta_2$, show that the differenced equation can be written as

$$\Delta\log(crime_i) = \delta_0 + \delta_1\Delta avgclr_i + \Delta u_i,$$

where $\delta_1 = 2\beta_1$ and $avgclri = (clrprc_{i,-1} + clrprc_{i,-2})/2$ is the average clear-up percentage over the previous two years.

(iii) Estimate the equation from part (ii). Compare the adjusted R-squared with that in (13.22). Which model would you finally use?

C7 Use GPA3 for this exercise. The data set is for 366 student-athletes from a large university for fall and spring semesters. [A similar analysis is in Maloney and McCormick (1993), but here we use a true panel data set.] Because you have two terms of data for each student, an unobserved effects model is appropriate. The primary question of interest is this: Do athletes perform more poorly in school during the semester their sport is in season?

(i) Use pooled OLS to estimate a model with term GPA (*trmgpa*) as the dependent variable. The explanatory variables are *spring*, *sat*, *hsperc*, *female*, *black*, *white*, *frstsem*, *tothrs*, *crsgpa*, and *season*. Interpret the coefficient on *season*. Is it statistically significant?

(ii) Most of the athletes who play their sport only in the fall are football players. Suppose the ability levels of football players differ systematically from those of other athletes. If ability is not adequately captured by SAT score and high school percentile, explain why the pooled OLS estimators will be biased.

(iii) Now, use the data differenced across the two terms. Which variables drop out? Now, test for an in-season effect.

(iv) Can you think of one or more potentially important, time-varying variables that have been omitted from the analysis?

C8 VOTE2 includes panel data on House of Representatives elections in 1988 and 1990. Only winners from 1988 who are also running in 1990 appear in the sample; these are the incumbents. An unobserved effects model explaining the share of the incumbent's vote in terms of expenditures by both candidates is

$$vote_{it} = \beta_0 + \delta_0 d90_t + \beta_1\log(inexp_{it}) + \beta_2\log(chexp_{it}) + \beta_3 incshr_{it} + a_i + u_{it},$$

where $incshr_{it}$ is the incumbent's share of total campaign spending (in percentage form). The unobserved effect a_i contains characteristics of the incumbent—such as "quality"—as well as things about the district that are constant. The incumbent's gender and party are constant over time, so these are subsumed in a_i. We are interested in the effect of campaign expenditures on election outcomes.

(i) Difference the given equation across the two years and estimate the differenced equation by OLS. Which variables are individually significant at the 5% level against a two-sided alternative?

(ii) In the equation from part (i), test for joint significance of $\Delta\log(inexp)$ and $\Delta\log(chexp)$. Report the p-value.

(iii) Reestimate the equation from part (i) using $\Delta incshr$ as the only independent variable. Interpret the coefficient on $\Delta incshr$. For example, if the incumbent's share of spending increases by 10 percentage points, how is this predicted to affect the incumbent's share of the vote?

(iv) Redo part (iii), but now use only the pairs that have repeat challengers. [This allows us to control for characteristics of the challengers as well, which would be in a_i. Levitt (1994) conducts a much more extensive analysis.]

C9 Use CRIME4 for this exercise.

(i) Add the logs of each wage variable in the data set and estimate the model by first differencing. How does including these variables affect the coefficients on the criminal justice variables in Example 13.9?

(ii) Do the wage variables in (i) all have the expected sign? Are they jointly significant? Explain.

C10 For this exercise, we use JTRAIN to determine the effect of the job training grant on hours of job training per employee. The basic model for the three years is

$$hrsemp_{it} = \beta_0 + \delta_1 d88_t + \delta_2 d89_t + \beta_1 grant_{it} + \beta_2 grant_{i,t-1} + \beta_3 \log(employ_{it}) + a_i + u_{it}.$$

(i) Estimate the equation using first differencing. How many firms are used in the estimation? How many total observations would be used if each firm had data on all variables (in particular, *hrsemp*) for the three time periods?

(ii) Interpret the coefficient on *grant* and comment on its significance.

(iii) Is it surprising that $grant_{-1}$ is insignificant? Explain.

(iv) Do larger firms train their employees more or less, on average? How big are the differences in training?

C11 The file MATHPNL contains panel data on school districts in Michigan for the years 1992 through 1998. It is the district-level analogue of the school-level data used by Papke (2005). The response variable of interest in this question is *math4*, the percentage of fourth graders in a district receiving a passing score on a standardized math test. The key explanatory variable is *rexpp*, which is real expenditures per pupil in the district. The amounts are in 1997 dollars. The spending variable will appear in logarithmic form.

(i) Consider the static unobserved effects model

$$math4_{it} = \delta_1 y93_t + \cdots + \delta_6 y98_t + \beta_1 \log(rexpp_{it}) + \beta_2 \log(enrol_{it}) + \beta_3 lunch_{it} + a_i + u_{it},$$

where $enrol_{it}$ is total district enrollment and $lunch_{it}$ is the percentage of students in the district eligible for the school lunch program. (So $lunch_{it}$ is a pretty good measure of the district-wide poverty rate.) Argue that $\beta_1/10$ is the percentage point change in $math4_{it}$ when real per-student spending increases by roughly 10%.

(ii) Use first differencing to estimate the model in part (i). The simplest approach is to allow an intercept in the first-differenced equation and to include dummy variables for the years 1994 through 1998. Interpret the coefficient on the spending variable.

(iii) Now, add one lag of the spending variable to the model and reestimate using first differencing. Note that you lose another year of data, so you are only using changes starting in 1994. Discuss the coefficients and significance on the current and lagged spending variables.

(iv) Obtain heteroskedasticity-robust standard errors for the first-differenced regression in part (iii). How do these standard errors compare with those from part (iii) for the spending variables?

(v) Now, obtain standard errors robust to both heteroskedasticity and serial correlation. What does this do to the significance of the lagged spending variable?

(vi) Verify that the differenced errors $r_{it} = \Delta u_{it}$ have negative serial correlation by carrying out a test of AR(1) serial correlation.

(vii) Based on a fully robust joint test, does it appear necessary to include the enrollment and lunch variables in the model?

C12 Use the data in MURDER for this exercise.

(i) Using the years 1990 and 1993, estimate the equation

$$mrdrte_{it} = \delta_0 + \delta_1 d93_t + \beta_1 exec_{it} + \beta_2 unem_{it} + a_i + u_{it}, t = 1, 2$$

by pooled OLS and report the results in the usual form. Do not worry that the usual OLS standard errors are inappropriate because of the presence of a_i. Do you estimate a deterrent effect of capital punishment?

(ii) Compute the FD estimates (use only the differences from 1990 to 1993; you should have 51 observations in the FD regression). Now what do you conclude about a deterrent effect?

(iii) In the FD regression from part (ii), obtain the residuals, say, $\hat{e}_i$. Run the Breusch-Pagan regression $\hat{e}_i^2$ on $\Delta exec_i$, $\Delta unem_i$ and compute the F test for heteroskedasticity. Do the same for the special case of the White test [that is, regress $\hat{e}_i^2$ on $\hat{y}_i$, $\hat{y}_i^2$, where the fitted values are from part (ii)]. What do you conclude about heteroskedasticity in the FD equation?

(iv) Run the same regression from part (ii), but obtain the heteroskedasticity-robust t statistics. What happens?

(v) Which t statistic on $\Delta exec_i$ do you feel more comfortable relying on, the usual one or the heteroskedasticity-robust one? Why?

C13 Use the data in WAGEPAN for this exercise.

(i) Consider the unobserved effects model

$$\begin{aligned} lwage_{it} = {} & \beta_0 + \delta_1 d81_t + \cdots + \delta_7 d87_t + \beta_1 educ_i \\ & + \gamma_1 d81_t educ_i + \cdots + \delta_7 d87_t educ_i + \beta_2 union_{it} + a_i + u_{it}, \end{aligned}$$

where a_i is allowed to be correlated with $educ_i$ and $union_{it}$. Which parameters can you estimate using first differencing?

(ii) Estimate the equation from part (i) by FD, and test the null hypothesis that the return to education has not changed over time.

(iii) Test the hypothesis from part (ii) using a fully robust test, that is, one that allows arbitrary heteroskedasticity and serial correlation in the FD errors, Δu_{it}. Does your conclusion change?

(iv) Now allow the union differential to change over time (along with education) and estimate the equation by FD. What is the estimated union differential in 1980? What about 1987? Is the difference statistically significant?

(v) Test the null hypothesis that the union differential has not changed over time, and discuss your results in light of your answer to part (iv).

C14 Use the data in JTRAIN3 for this exercise.

(i) Estimate the simple regression model $re78 = \beta_0 + \beta_1 train + u$, and report the results in the usual form. Based on this regression, does it appear that job training, which took place in 1976 and 1977, had a positive effect on real labor earnings in 1978?

(ii) Now use the change in real labor earnings, $cre = re78 - re75$, as the dependent variable. (We need not difference *train* because we assume there was no job training prior to 1975. That is, if we define $ctrain = train78 - train75$ then $ctrain = train78$ because $train75 = 0$.) Now what is the estimated effect of training? Discuss how it compares with the estimate in part (i).

(iii) Find the 95% confidence interval for the training effect using the usual OLS standard error and the heteroskedasticity-robust standard error, and describe your findings.

C15 The data set HAPPINESS contains independently pooled cross sections for the even years from 1994 through 2006, obtained from the General Social Survey. The dependent variable for this problem is a measure of "happiness," *vhappy*, which is a binary variable equal to one if the person reports being "very happy" (as opposed to just "pretty happy" or "not too happy").

(i) Which year has the largest number of observations? Which has the smallest? What is the percentage of people in the sample reporting they are "very happy"?

(ii) Regress *vhappy* on all of the year dummies, leaving out *y*94 so that 1994 is the base year. Compute a heteroskedasticity-robust statistic of the null hypothesis that the proportion of very happy people has not changed over time. What is the *p*-value of the test?

(iii) To the regression in part (ii), add the dummy variables *occattend* and *regattend*. Interpret their coefficients. (Remember, the coefficients are interpreted relative to a base group.) How would you summarize the effects of church attendance on happiness?

(iv) Define a variable, say *highinc*, equal to one if family income is above $25,000. (Unfortunately, the same threshold is used in each year, and so inflation is not accounted for. Also, $25,000 is hardly what one would consider "high income.") Include *highinc*, *unem10*, *educ*, and *teens* in the regression in part (iii). Is the coefficient on *regattend* affected much? What about its statistical significance?

(v) Discuss the signs, magnitudes, and statistical significance of the four new variables in part (iv). Do the estimates make sense?

(vi) Controlling for the factors in part (iv), do there appear to be differences in happiness by gender or race? Justify your answer.

C16 Use the data in COUNTYMURDERS for this exercise. The data set covers murders and executions (capital punishment) for 2,197 counties in the United States.

(i) Find the average value of *murdrate* across all counties and years. What is the standard deviation? For what percentage of the sample is *murdrate* equal to zero?

(ii) How many observations have *execs* equal to zero? What is the maximum value of *execs*? Why is the average of *execs* so small?

(iii) Consider the model

$$murdrate_{it} = \theta_t + \beta_1 execs_{it} + \beta_2 execs_{i,t-1} + \beta_3 percblack_{it} + \beta_4 percmale_i + \beta_5 perc1019 + \beta_6 perc2029 + a_i + u_{it},$$

where θ_t represents a different intercept for each time period, a_i is the county fixed effect, and u_{it} is the idiosyncratic error. What do we need to assume about a_i and the execution variables in order for pooled OLS to consistently estimate the parameters, in particular, β_1 and β_2?

(iv) Apply OLS to the equation from part (ii) and report the estimates of β_1 and β_2, along with the usual pooled OLS standard errors. Do you estimate that executions have a deterrent effect on murders? What do you think is happening?

(v) Even if the pooled OLS estimators are consistent, do you trust the standard errors obtained from part (iv)? Explain.

(vi) Now estimate the equation in part (iii) using first differencing to remove a_i. What are the new estimates of β_1 and β_2? Are they very different from the estimates from part (iv)?

(vii) Using the estimates from part (vi), can you say there is evidence of a statistically significant deterrent effect of capital punishment on the murder rate? If possible, in addition to the usual OLS standard errors, use those that are robust to any kind of serial correlation or heteroskedasticity in the FD errors.

APPENDIX 13A

13A.1 Assumptions for Pooled OLS Using First Differences

In this appendix, we provide careful statements of the assumptions for the first-differencing estimator. Verification of the following claims can be found in Wooldridge (2010, Chapter 10).

Assumption FD.1

For each i, the model is

$$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \; t = 1, \ldots, T,$$

where the β_j are the parameters to estimate and a_i is the unobserved effect.

Assumption FD.2

We have a random sample from the cross section.

Assumption FD.3

Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables.

For the next assumption, it is useful to let $\mathbf{X}_i$ denote the explanatory variables for all time periods for cross-sectional observation i; thus, $\mathbf{X}_i$ contains x_{itj}, $t = 1, \ldots, T, j = 1, \ldots, k$.

Assumption FD.4

For each t, the expected value of the idiosyncratic error given the explanatory variables in *all* time periods and the unobserved effect is zero: $\mathrm{E}(u_{it}|\mathbf{X}_i, a_i) = 0$.

When Assumption FD.4 holds, we sometimes say that the x_{itj} are *strictly exogenous conditional on the unobserved effect*. The idea is that, once we control for a_i, there is no correlation between the x_{isj} and the remaining idiosyncratic error, u_{it}, for all s and t.

As stated, Assumption FD.4 is stronger than necessary. We use this form of the assumption because it emphasizes that we are interested in the equation

$$\mathrm{E}(y_{it}|\mathbf{X}_i, a_i) = \mathrm{E}(y_{it}|\mathbf{x}_{it}, a_i) = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i,$$

so that the β_j measure partial effects of the observed explanatory variables holding fixed, or "controlling for," the unobserved effect, a_i. Nevertheless, an important implication of FD.4, and one that is sufficient for the unbiasedness of the FD estimator, is $\mathrm{E}(\Delta u_{it}|\mathbf{X}_i) = 0, t = 2, \ldots, T$. In fact, for consistency we can simply assume that Δx_{itj} is uncorrelated with Δu_{it} for all $t = 2, \ldots, T$ and $j = 1, \ldots, k$. See Wooldridge (2010, Chapter 10) for further discussion.

Under these first four assumptions, the first-difference estimators are unbiased. The key assumption is FD.4, which is strict exogeneity of the explanatory variables. Under these same assumptions, we can also show that the FD estimator is consistent with a fixed T and as $N \to \infty$ (and perhaps more generally).

The next two assumptions ensure that the standard errors and test statistics resulting from pooled OLS on the first differences are (asymptotically) valid.

Assumption FD.5

The variance of the differenced errors, conditional on all explanatory variables, is constant: $\text{Var}(\Delta u_{it}|\mathbf{X}_i) = \sigma^2, t = 2, \ldots, T$.

Assumption FD.6

For all $t \neq s$, the *differences* in the idiosyncratic errors are uncorrelated (conditional on all explanatory variables): $\text{Cov}(\Delta u_{it}, \Delta u_{is}|\mathbf{X}_i) = 0, t \neq s$.

Assumption FD.5 ensures that the differenced errors, Δu_{it}, are homoskedastic. Assumption FD.6 states that the differenced errors are serially uncorrelated, which means that the u_{it} follow a random walk across time (see Chapter 11). Under Assumptions FD.1 through FD.6, the FD estimator of the β_j is the best linear unbiased estimator (conditional on the explanatory variables).

Assumption FD.7

Conditional on $\mathbf{X}_i$, the Δu_{it} are independent and identically distributed normal random variables.

When we add Assumption FD.7, the FD estimators are normally distributed, and the t and F statistics from pooled OLS on the differences have exact t and F distributions. Without FD.7, we can rely on the usual asymptotic approximations.

13A.2 Computing Standard Errors Robust to Serial Correlation and Heteroskedasticity of Unknown Form

Because the FD estimator is consistent as $N \to \infty$ under Assumptions FD.1 through FD.4, it would be very handy to have a simple method of obtaining proper standard errors and test statistics that allow for any kind of serial correlation or heteroskedasticity in the FD errors, $e_{it} = \Delta u_{it}$. Fortunately, provided N is moderately large, and T is not "too large," fully robust standard errors and test statistics are readily available. As mentioned in the text, a detailed treatment is above the level of this text. The technical arguments combine the insights described in Chapters 8 and 12, where statistics robust to heteroskedasticity and serial correlation are discussed. Actually, there is one important advantage with panel data: because we have a (large) cross section, we can allow unrestricted serial correlation in the errors $\{e_{it}\}$, provided T is not too large. We can contrast this situation with the Newey-West approach in Section 12-5, where the estimated covariances must be downweighted as the observations get farther apart in time. Wooldridge (2010, Chapter 10) provides further discussion.

The general approach to obtaining fully robust standard errors and test statistics in the context of panel data is known as **clustering**, and ideas have been borrowed from the cluster sampling literature. The idea is that each cross-sectional unit is defined as a cluster of observations over time, and arbitrary correlation—serial correlation—and changing variances are allowed within each cluster. Because of the relationship to cluster sampling, many econometric software packages have options for clustering standard errors and test statistics. The resulting standard errors are often called cluster-robust standard errors. As a bonus, such standard errors are also robust to heteroskedasticity of unknown form.

Most commands look something like

regress cy cd2 cd3 . . . cdT cx1 cx2 . . . cxk, noconstant cluster(id),

where "id" is a variable containing unique identifiers for the cross-sectional units and the "c" before each variable denotes "change." The option "cluster(id)" at the end of the "regress" command

tells the software to report all standard errors and test statistics—including t statistics and F-type statistics—so that they are valid, in large cross sections, with any kind of serial correlation or heteroskedasticity. The "noconstant" option suppresses the intercept, as it gets eliminated via the differencing. An alternative is to allow a constant and to include the time dummies $d3$, $d4$, . . . , dT in levels form. This will not change the estimates on the explanatory variables of interest, just on the time effects.

Some packages have an option that does not require differencing ahead of time, which saves some work, is likely to result in fewer mistakes, and also reminds us that the equation of interest is in levels, and differencing results in an estimating equation:

```
regress D.(y d2 d3 . . . dT x1 x2 . . . xk), noconstant cluster(id)
```

where "D.()" denotes differencing everything in parentheses.

Reporting cluster-robust standard errors and test statistics is now very common in modern empirical work with panel data. Often the standard errors will be larger than either the usual OLS standard errors or those that correct only for heteroskedasticity, but it is possible for cluster-robust standard errors to be smaller, too. In any case, provided N is moderately large and T is not too large, the cluster-robust standard errors better reflect the uncertainty in the pooled OLS coefficients.

There is one important point about clustering to account for serial correlation: it does not account for any cross-sectional correlation. In fact, we assume that the draws of units i from the population are independent. Removing one potential source of cross-sectional correlation, the unobserved effect a_i, can help. Also, controlling for aggregate time effects through the time dummies accounts for cross-sectional correlation caused by common shocks.

Math Refresher A

Basic Mathematical Tools

This Math Refresher covers some basic mathematics that are used in econometric analysis. We summarize various properties of the summation operator, study properties of linear and certain nonlinear equations, and review proportions and percentages. We also present some special functions that often arise in applied econometrics, including quadratic functions and the natural logarithm. The first four sections require only basic algebra skills. Section A-5 contains a brief review of differential calculus; although a knowledge of calculus is not necessary to understand most of the text, it is used in some end-of-chapter appendices and in several of the more advanced chapters in Part 3.

A-1 The Summation Operator and Descriptive Statistics

The **summation operator** is a useful shorthand for manipulating expressions involving the sums of many numbers, and it plays a key role in statistics and econometric analysis. If $\{x_i: i = 1, \ldots, n\}$ denotes a sequence of n numbers, then we write the sum of these numbers as

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \cdots + x_n. \qquad [A.1]$$

With this definition, the summation operator is easily shown to have the following properties:

Property Sum.1: For any constant c,

$$\sum_{i=1}^{n} c = nc. \qquad [A.2]$$

Property Sum.2: For any constant c,

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i. \qquad [A.3]$$

Property Sum.3: If $\{(x_i, y_i): i = 1, 2, \ldots, n\}$ is a set of n pairs of numbers, and a and b are constants, then

$$\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n}x_i + b\sum_{i=1}^{n}y_i. \qquad \text{[A.4]}$$

It is also important to be aware of some things that *cannot* be done with the summation operator. Let$\{(x_i, y_i): i = 1, 2, \ldots, n\}$ again be a set of n pairs of numbers with $y_i \neq 0$ for each i. Then,

$$\sum_{i=1}^{n}(x_i/y_i) \neq \left(\sum_{i=1}^{n}x_i\right)\Big/\left(\sum_{i=1}^{n}y_i\right).$$

In other words, the sum of the ratios is not the ratio of the sums. In the $n = 2$ case, the application of familiar elementary algebra also reveals this lack of equality: $x_1/y_1 + x_2/y_2 \neq (x_1 + x_2)/(y_1 + y_2)$. Similarly, the sum of the squares is not the square of the sum: $\sum_{i=1}^{n}x_i^2 \neq (\sum_{i=1}^{n}x_i)^2$, except in special cases. That these two quantities are not generally equal is easiest to see when $n = 2$: $x_1^2 + x_2^2 \neq (x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$.

Given n numbers $\{x_i: i = 1, \ldots, n\}$, we compute their **average** or *mean* by adding them up and dividing by n:

$$\bar{x} = (1/n)\sum_{i=1}^{n}x_i. \qquad \text{[A.5]}$$

When the x_i are a sample of data on a particular variable (such as years of education), we often call this the *sample average* (or *sample mean*) to emphasize that it is computed from a particular set of data. The sample average is an example of a **descriptive statistic**; in this case, the statistic describes the central tendency of the set of points x_i.

There are some basic properties about averages that are important to understand. First, suppose we take each observation on x and subtract off the average: $d_i \equiv x_i - \bar{x}$ (the "d" here stands for *deviation* from the average). Then, the sum of these deviations is always zero:

$$\sum_{i=1}^{n}d_i = \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x} = \sum_{i=1}^{n}x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

We summarize this as

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0. \qquad \text{[A.6]}$$

A simple numerical example shows how this works. Suppose $n = 5$ and $x_1 = 6$, $x_2 = 1$, $x_3 = -2$, $x_4 = 0$, and $x_5 = 5$. Then, $\bar{x} = 2$, and the demeaned sample is $\{4, -1, -4, -2, 3\}$. Adding these gives zero, which is just what equation (A.6) says.

In our treatment of regression analysis in Chapter 2, we need to know some additional algebraic facts involving deviations from sample averages. An important one is that the sum of squared deviations is the sum of the squared x_i minus n times the square of $\bar{x}$:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n(\bar{x})^2. \qquad \text{[A.7]}$$

This can be shown using basic properties of the summation operator:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$
$$= \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n(\bar{x})^2$$
$$= \sum_{i=1}^{n}x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2 = \sum_{i=1}^{n}x_i^2 - n(\bar{x})^2.$$

Given a data set on two variables, $\{(x_i, y_i): i = 1, 2, \ldots, n\}$, it can also be shown that

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}x_i(y_i - \bar{y}) \qquad \text{[A.8]}$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n}x_iy_i - n(\bar{x}\cdot\bar{y});$$

this is a generalization of equation (A.7). (There, $y_i = x_i$ for all i.)

The average is the measure of central tendency that we will focus on in most of this text. However, it is sometimes informative to use the **median** (or *sample median*) to describe the central value. To obtain the median of the n numbers $\{x_1, \ldots, x_n\}$, we first order the values of the x_i from smallest to largest. Then, if n is odd, the sample median is the middle number of the ordered observations. For example, given the numbers $\{-4, 8, 2, 0, 21, -10, 18\}$, the median value is 2 (because the ordered sequence is $\{-10, -4, 0, 2, 8, 18, 21\}$). If we change the largest number in this list, 21, to twice its value, 42, the median is still 2. By contrast, the sample average would increase from 5 to 8, a sizable change. Generally, the median is less sensitive than the average to changes in the extreme values (large or small) in a list of numbers. This is why "median incomes" or "median housing values" are often reported, rather than averages, when summarizing income or housing values in a city or county.

If n is even, there is no unique way to define the median because there are two numbers at the center. Usually, the median is defined to be the average of the two middle values (again, after ordering the numbers from smallest to largest). Using this rule, the median for the set of numbers $\{4, 12, 2, 6\}$ would be $(4 + 6)/2 = 5$.

A-2 Properties of Linear Functions

Linear functions play an important role in econometrics because they are simple to interpret and manipulate. If x and y are two variables related by

$$y = \beta_0 + \beta_1 x, \qquad \text{[A.9]}$$

then we say that y is a **linear function** of x, and β_0 and β_1 are two parameters (numbers) describing this relationship. The **intercept** is β_0, and the **slope** is β_1.

The defining feature of a linear function is that the change in y is always β_1 times the change in x:

$$\Delta y = \beta_1 \Delta x, \qquad \text{[A.10]}$$

where Δ denotes "change." In other words, the **marginal effect** of x on y is constant and equal to β_1.

EXAMPLE A.1 **Linear Housing Expenditure Function**

Suppose that the relationship between monthly housing expenditure and monthly income is

$$housing = 164 + .27\, income. \qquad [A.11]$$

Then, for each additional dollar of income, 27 cents is spent on housing. If family income increases by \$200, then housing expenditure increases by $(.27)200 = \$54$. This function is graphed in Figure A.1.

According to equation (A.11), a family with no income spends \$164 on housing, which of course cannot be literally true. For low levels of income, this linear function would not describe the relationship between *housing* and *income* very well, which is why we will eventually have to use other types of functions to describe such relationships.

In (A.11), the *marginal propensity to consume* (MPC) housing out of income is .27. This is different from the *average propensity to consume* (APC), which is

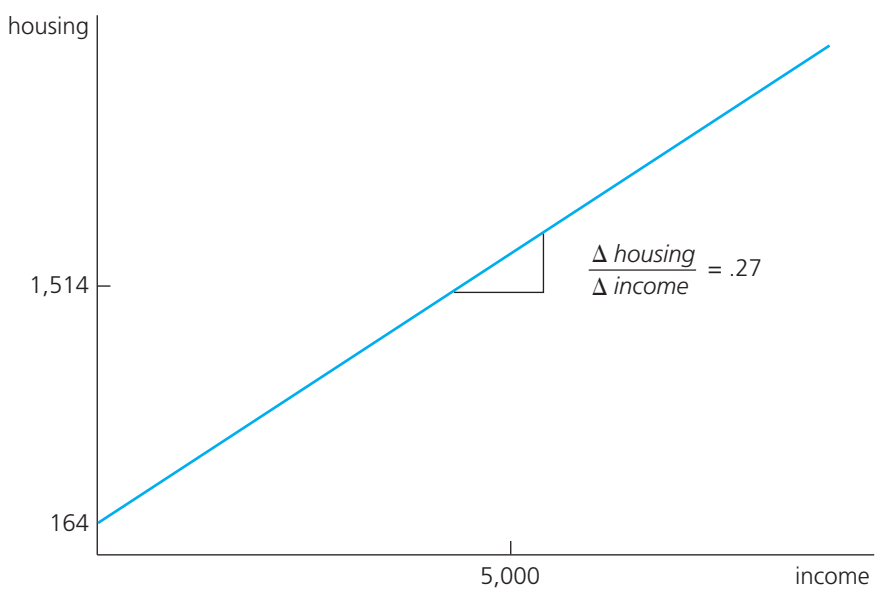
$$\frac{housing}{income} = 164/income + .27.$$

The APC is not constant; it is always larger than the MPC, and it gets closer to the MPC as income increases.

Linear functions are easily defined for more than two variables. Suppose that y is related to two variables, x_1 and x_2, in the general form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \qquad [A.12]$$

FIGURE A.1 Graph of *housing* = 164 + .27 *income*.



It is rather difficult to envision this function because its graph is three-dimensional. Nevertheless, β_0 is still the intercept (the value of y when $x_1 = 0$ and $x_2 = 0$), and β_1 and β_2 measure particular slopes. From (A.12), the change in y, for given changes in x_1 and x_2, is

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2. \qquad [A.13]$$

If x_2 does not change, that is, $\Delta x_2 = 0$, then we have

$$\Delta y = \beta_1 \Delta x_1 \text{ if } \Delta x_2 = 0,$$

so that β_1 is the slope of the relationship in the direction of x_1:

$$\beta_1 = \frac{\Delta y}{\Delta x_1} \text{ if } \Delta x_2 = 0.$$

Because it measures how y changes with x_1, holding x_2 fixed, β_1 is often called the **partial effect** of x_1 on y. Because the partial effect involves holding other factors fixed, it is closely linked to the notion of **ceteris paribus**. The parameter β_2 has a similar interpretation: $\beta_2 = \Delta y/\Delta x_2$ if $\Delta x_1 = 0$, so that β_2 is the partial effect of x_2 on y.

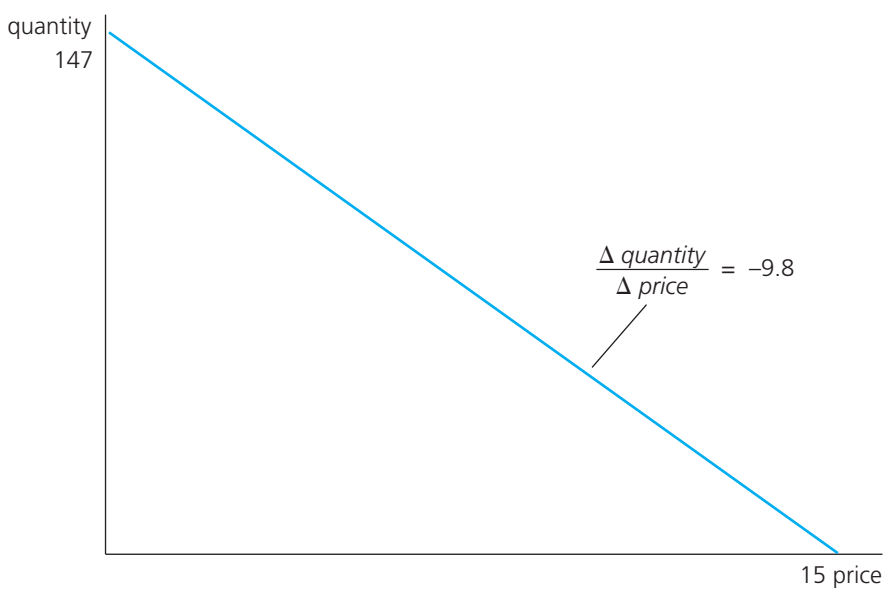
EXAMPLE A.2 Demand for Compact Discs

For college students, suppose that the monthly quantity demanded of compact discs is related to the price of compact discs and monthly discretionary income by

$$quantity = 120 - 9.8\, price + .03\, income,$$

where *price* is dollars per disc and *income* is measured in dollars. The *demand curve* is the relationship between *quantity* and *price*, holding *income* (and other factors) fixed. This is graphed in two dimensions in Figure A.2 at an income level of $900. The slope of the demand curve, -9.8, is the *partial effect* of price on quantity: holding income fixed, if the price of compact discs increases by one dollar, then the quantity demanded falls by 9.8. (We abstract from the fact that CDs can only be purchased in discrete units.) An increase in income simply shifts the demand curve up (changes the intercept), but the slope remains the same.

FIGURE A.2 Graph of *quantity* = 120 − 9.8 *price* + .03 *income*, with income fixed at $900.



A-3 Proportions and Percentages

Proportions and percentages play such an important role in applied economics that it is necessary to become very comfortable in working with them. Many quantities reported in the popular press are in the form of percentages; a few examples are interest rates, unemployment rates, and high school graduation rates.

An important skill is being able to convert proportions to percentages and vice versa. A percentage is easily obtained by multiplying a proportion by 100. For example, if the proportion of adults in a county with a high school degree is .82, then we say that 82% (82 percent) of adults have a high school degree. Another way to think of percentages and proportions is that a proportion is the decimal form of a percentage. For example, if the marginal tax rate for a family earning $30,000 per year is reported as 28%, then the proportion of the next dollar of income that is paid in income taxes is .28 (or 28¢).

When using percentages, we often need to convert them to decimal form. For example, if a state sales tax is 6% and $200 is spent on a taxable item, then the sales tax paid is 200(.06) = $12. If the annual return on a certificate of deposit (CD) is 7.6% and we invest $3,000 in such a CD at the beginning of the year, then our interest income is 3,000(.076) = $228. As much as we would like it, the interest income is not obtained by multiplying 3,000 by 7.6.

We must be wary of proportions that are sometimes incorrectly reported as percentages in the popular media. If we read, "The percentage of high school students who drink alcohol is .57," we know that this really means 57% (not just over one-half of a percent, as the statement literally implies). College volleyball fans are probably familiar with press clips containing statements such as "Her hitting percentage was .372." This really means that her hitting percentage was 37.2%.

In econometrics, we are often interested in measuring the *changes* in various quantities. Let x denote some variable, such as an individual's income, the number of crimes committed in a community, or the profits of a firm. Let x_0 and x_1 denote two values for x: x_0 is the initial value, and x_1 is the subsequent value. For example, x_0 could be the annual income of an individual in 1994 and x_1 the income of the same individual in 1995. The **proportionate change** in x in moving from x_0 to x_1, sometimes called the **relative change,** is simply

$$(x_1 - x_0)/x_0 = \Delta x/x_0, \qquad \text{[A.14]}$$

assuming, of course, that $x_0 \neq 0$. In other words, to get the proportionate change, we simply divide the change in x by its initial value. This is a way of standardizing the change so that it is free of units. For example, if an individual's income goes from $30,000 per year to $36,000 per year, then the proportionate change is 6,000/30,000 = .20.

It is more common to state changes in terms of percentages. The **percentage change** in x in going from x_0 to x_1 is simply 100 times the proportionate change:

$$\%\Delta x = 100(\Delta x/x_0); \qquad \text{[A.15]}$$

the notation "$\%\Delta x$" is read as "the percentage change in x." For example, when income goes from $30,000 to $33,750, income has increased by 12.5%; to get this, we simply multiply the proportionate change, .125, by 100.

Again, we must be on guard for proportionate changes that are reported as percentage changes. In the previous example, for instance, reporting the percentage change in income as .125 is incorrect and could lead to confusion.

When we look at changes in things like dollar amounts or population, there is no ambiguity about what is meant by a percentage change. By contrast, interpreting percentage change calculations can be tricky when the variable of interest is itself a percentage, something that happens often in economics and other social sciences. To illustrate, let x denote the percentage of adults in a particular city having a college education. Suppose the initial value is $x_0 = 24$ (24% have a college education), and the new

value is $x_1 = 30$. We can compute two quantities to describe how the percentage of college-educated people has changed. The first is the change in x, Δx. In this case, $\Delta x = x_1 - x_0 = 6$: the percentage of people with a college education has increased by six *percentage points*. On the other hand, we can compute the percentage change in x using equation (A.15): $\%\Delta x = 100[(30 - 24)/24] = 25$.

In this example, the percentage point change and the percentage change are very different. The **percentage point change** is just the change in the percentages. The percentage change is the change relative to the initial value. Generally, we must pay close attention to which number is being computed. The careful researcher makes this distinction perfectly clear; unfortunately, in the popular press as well as in academic research, the type of reported change is often unclear.

EXAMPLE A.3 **Michigan Sales Tax Increase**

In March 1994, Michigan voters approved a sales tax increase from 4% to 6%. In political advertisements, supporters of the measure referred to this as a two percentage point increase, or an increase of two cents on the dollar. Opponents of the tax increase called it a 50% increase in the sales tax rate. Both claims are correct; they are simply different ways of measuring the increase in the sales tax. Naturally, each group reported the measure that made its position most favorable.

For a variable such as salary, it makes no sense to talk of a "percentage point change in salary" because salary is not measured as a percentage. We can describe a change in salary either in dollar or percentage terms.

A-4 Some Special Functions and Their Properties

In Section A-2, we reviewed the basic properties of linear functions. We already indicated one important feature of functions like $y = \beta_0 + \beta_1 x$: a one-unit change in x results in the *same* change in y, regardless of the initial value of x. As we noted earlier, this is the same as saying the marginal effect of x on y is constant, something that is not realistic for many economic relationships. For example, the important economic notion of *diminishing marginal returns* is not consistent with a linear relationship.

In order to model a variety of economic phenomena, we need to study several nonlinear functions. A **nonlinear function** is characterized by the fact that the change in y for a given change in x depends on the starting value of x. Certain nonlinear functions appear frequently in empirical economics, so it is important to know how to interpret them. A complete understanding of nonlinear functions takes us into the realm of calculus. Here, we simply summarize the most significant aspects of the functions, leaving the details of some derivations for Section A-5.

A-4a Quadratic Functions

One simple way to capture diminishing returns is to add a quadratic term to a linear relationship. Consider the equation

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \qquad [A.16]$$

where β_0, β_1, and β_2 are parameters. When $\beta_1 > 0$ and $\beta_2 < 0$, the relationship between y and x has the parabolic shape given in Figure A.3, where $\beta_0 = 6$, $\beta_1 = 8$, and $\beta_2 = -2$.

When $\beta_1 > 0$ and $\beta_2 < 0$, it can be shown (using calculus in the next section) that the *maximum* of the function occurs at the point

$$x^* = \beta_1/(-2\beta_2). \qquad \text{[A.17]}$$

For example, if $y = 6 + 8x - 2x^2$ (so $\beta_1 = 8$ and $\beta_2 = -2$), then the largest value of y occurs at $x^* = 8/4 = 2$, and this value is $6 + 8(2) - 2(2)^2 = 14$ (see Figure A.3).

The fact that equation (A.16) implies a **diminishing marginal effect** of x on y is easily seen from its graph. Suppose we start at a low value of x and then increase x by some amount, say, c. This has a larger effect on y than if we start at a higher value of x and increase x by the same amount c. In fact, once $x > x^*$, an increase in x actually decreases y.

The statement that x has a diminishing marginal effect on y is the same as saying that the slope of the function in Figure A.3 decreases as x increases. Although this is clear from looking at the graph, we usually want to quantify how quickly the slope is changing. An application of calculus gives the approximate slope of the quadratic function as

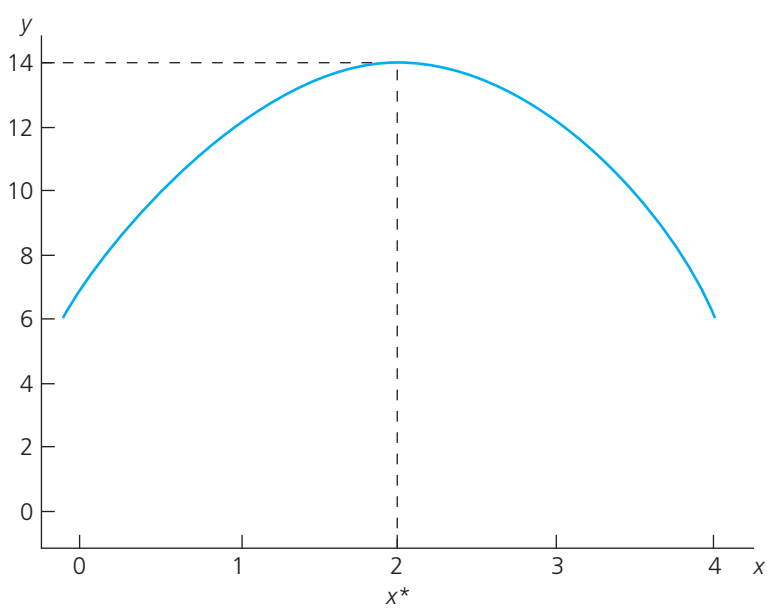
$$slope = \frac{\Delta y}{\Delta x} \approx \beta_1 + 2\beta_2 x, \qquad \text{[A.18]}$$

for "small" changes in x. [The right-hand side of equation (A.18) is the **derivative** of the function in equation (A.16) with respect to x.] Another way to write this is

$$\Delta y \approx (\beta_1 + 2\beta_2 x)\Delta x \text{ for "small" } \Delta x. \qquad \text{[A.19]}$$

To see how well this approximation works, consider again the function $y = 6 + 8x - 2x^2$. Then, according to equation (A.19), $\Delta y \approx (8 - 4x)\Delta x$. Now, suppose we start at $x = 1$ and change x by $\Delta x = .1$. Using (A.19), $\Delta y \approx (8 - 4)(.1) = .4$. Of course, we can compute the change exactly by finding the values of y when $x = 1$ and $x = 1.1$: $y_0 = 6 + 8(1) - 2(1)^2 = 12$ and $y_1 = 6 + 8(1.1) - 2(1.1)^2 = 12.38$, so the exact change in y is .38. The approximation is pretty close in this case.

FIGURE A.3 Graph of $y = 6 + 8x - 2x^2$.



Now, suppose we start at $x = 1$ but change x by a larger amount: $\Delta x = .5$. Then, the approximation gives $\Delta y \approx 4(.5) = 2$. The exact change is determined by finding the difference in y when $x = 1$ and $x = 1.5$. The former value of y was 12, and the latter value is $6 + 8(1.5) - 2(1.5)^2 = 13.5$, so the actual change is 1.5 (not 2). The approximation is worse in this case because the change in x is larger.

For many applications, equation (A.19) can be used to compute the approximate marginal effect of x on y for any initial value of x and small changes. And, we can always compute the exact change if necessary.

EXAMPLE A.4 A Quadratic Wage Function

Suppose the relationship between hourly wages and years in the workforce (*exper*) is given by

$$wage = 5.25 + .48\, exper - .008\, exper^2. \qquad [A.20]$$

This function has the same general shape as the one in Figure A.3. Using equation (A.17), *exper* has a positive effect on wage up to the turning point, $exper^* = .48/[2(.008)] = 30$. The first year of experience is worth approximately .48, or 48 cents [see (A.19) with $x = 0$, $\Delta x = 1$]. Each additional year of experience increases wage by less than the previous year—reflecting a diminishing marginal return to experience. At 30 years, an additional year of experience would actually lower the wage. This is not very realistic, but it is one of the consequences of using a quadratic function to capture a diminishing marginal effect: at some point, the function must reach a maximum and curve downward. For practical purposes, the point at which this happens is often large enough to be inconsequential, but not always.

The graph of the quadratic function in (A.16) has a U-shape if $\beta_1 < 0$ and $\beta_2 > 0$, in which case there is an increasing marginal return. The minimum of the function is at the point $-\beta_1/(2\beta_2)$.

A-4b The Natural Logarithm

The nonlinear function that plays the most important role in econometric analysis is the **natural logarithm**. In this text, we denote the natural logarithm, which we often refer to simply as the **log function**, as

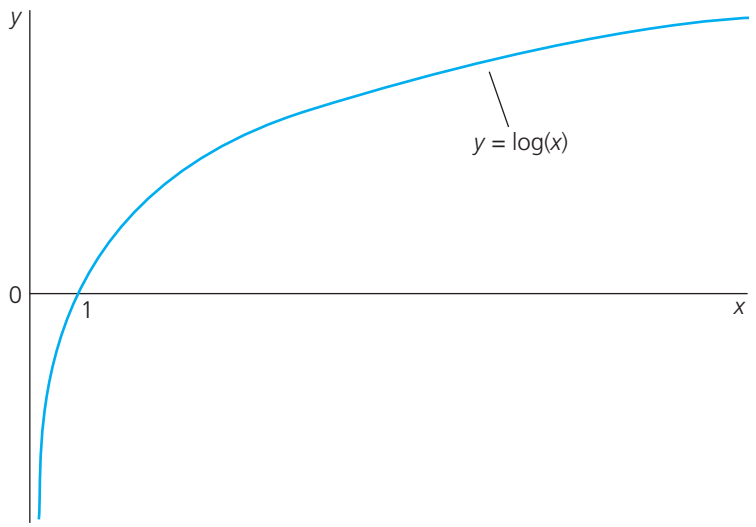
$$y = \log(x). \qquad [A.21]$$

You might remember learning different symbols for the natural log; $\ln(x)$ or $\log_e(x)$ are the most common. These different notations are useful when logarithms with several different bases are being used. For our purposes, only the natural logarithm is important, and so $\log(x)$ denotes the natural logarithm throughout this text. This corresponds to the notational usage in many statistical packages, although some use $\ln(x)$ [and most calculators use $\ln(x)$]. Economists use both $\log(x)$ and $\ln(x)$, which is useful to know when you are reading papers in applied economics.

The function $y = \log(x)$ is defined only for $x > 0$, and it is plotted in Figure A.4. It is not very important to know how the values of $\log(x)$ are obtained. For our purposes, the function can be thought of as a black box: we can plug in any $x > 0$ and obtain $\log(x)$ from a calculator or a computer.

Several things are apparent from Figure A.4. First, when $y = \log(x)$, the relationship between y and x displays diminishing marginal returns. One important difference between the log and the quadratic function in Figure A.3 is that when $y = \log(x)$, the effect of x on y never becomes negative: the

FIGURE A.4 Graph of $y = \log(x)$.



slope of the function gets closer and closer to zero as x gets large, but the slope never quite reaches zero and certainly never becomes negative.

The following are also apparent from Figure A.4:

$$\log(x) < 0 \text{ for } 0 < x < 1$$
$$\log(1) = 0$$
$$\log(x) > 0 \text{ for } x > 1.$$

In particular, $\log(x)$ can be positive or negative. Some useful algebraic facts about the log function are

$$\log(x_1 \cdot x_2) = \log(x_1) + \log(x_2), x_1, x_2 > 0$$
$$\log(x_1/x_2) = \log(x_1) - \log(x_2), x_1, x_2 > 0$$
$$\log(x^c) = c \log(x), x > 0, c \text{ any number.}$$

Occasionally, we will need to rely on these properties.

The logarithm can be used for various approximations that arise in econometric applications. First, $\log(1 + x) \approx x$ for $x \approx 0$. You can try this with $x = .02$, .1, and .5 to see how the quality of the approximation deteriorates as x gets larger. Even more useful is the fact that the difference in logs can be used to approximate proportionate changes. Let x_0 and x_1 be positive values. Then, it can be shown (using calculus) that

$$\log(x_1) - \log(x_0) \approx (x_1 - x_0)/x_0 = \Delta x/x_0 \qquad \text{[A.22]}$$

for small changes in x. If we multiply equation (A.22) by 100 and write $\Delta\log(x) = \log(x_1) - \log(x_0)$, then

$$100 \cdot \Delta\log(x) \approx \%\Delta x \qquad \text{[A.23]}$$

for small changes in x. The meaning of "small" depends on the context, and we will encounter several examples throughout this text.

Why should we approximate the percentage change using (A.23) when the exact percentage change is so easy to compute? Momentarily, we will see why the approximation in (A.23) is useful in econometrics. First, let us see how good the approximation is in two examples.

First, suppose $x_0 = 40$ and $x_1 = 41$. Then, the percentage change in x in moving from x_0 to x_1 is 2.5%, using $100(x_1 - x_0)/x_0$. Now, $\log(41) - \log(40) = .0247$ (to four decimal places), which when multiplied by 100 is very close to 2.5. The approximation works pretty well. Now, consider a much bigger change: $x_0 = 40$ and $x_1 = 60$. The exact percentage change is 50%. However, $\log(60) - \log(40) \approx .4055$, so the approximation gives 40.55%, which is much farther off.

Why is the approximation in (A.23) useful if it is only satisfactory for small changes? To build up to the answer, we first define the **elasticity** of y with respect to x as

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \frac{\%\Delta y}{\%\Delta x}. \qquad [A.24]$$

In other words, the elasticity of y with respect to x is the percentage change in y when x increases by 1%. This notion should be familiar from introductory economics.

If y is a linear function of x, $y = \beta_0 + \beta_1 x$, then the elasticity is

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{\beta_0 + \beta_1 x}, \qquad [A.25]$$

which clearly depends on the value of x. (This is a generalization of the well-known result from basic demand theory: the elasticity is not constant along a straight-line demand curve.)

Elasticities are of critical importance in many areas of applied economics, not just in demand theory. It is convenient in many situations to have *constant* elasticity models, and the log function allows us to specify such models. If we use the approximation in (A.23) for both x and y, then the elasticity is approximately equal to $\Delta\log(y)/\Delta\log(x)$. Thus, a constant elasticity model is approximated by the equation

$$\log(y) = \beta_0 + \beta_1 \log(x), \qquad [A.26]$$

and β_1 is the elasticity of y with respect to x (assuming that $x, y > 0$).

EXAMPLE A.5 Constant Elasticity Demand Function

If q is quantity demanded and p is price and these variables are related by

$$\log(q) = 4.7 - 1.25 \log(p),$$

then the price elasticity of demand is -1.25. Roughly, a 1% increase in price leads to a 1.25% fall in the quantity demanded.

For our purposes, the fact that β_1 in (A.26) is only close to the elasticity is not important. In fact, when the elasticity is defined using calculus—as in Section A-5—the definition is exact. For the purposes of econometric analysis, (A.26) defines a **constant elasticity model**. Such models play a large role in empirical economics.

Other possibilities for using the log function often arise in empirical work. Suppose that $y > 0$ and

$$\log(y) = \beta_0 + \beta_1 x. \qquad [A.27]$$

Then, $\Delta\log(y) = \beta_1\Delta x$, so $100\cdot\Delta\log(y) = (100\cdot\beta_1)\Delta x$. It follows that, when y and x are related by equation (A.27),

$$\%\Delta y \approx (100 \cdot \beta_1)\Delta x. \qquad [A.28]$$

EXAMPLE A.6 **Logarithmic Wage Equation**

Suppose that hourly wage and years of education are related by

$$\log(wage) = 2.78 + .094\, educ.$$

Then, using equation (A.28),

$$\%\Delta wage \approx 100(.094)\Delta educ = 9.4\ \Delta educ.$$

It follows that one more year of education increases hourly wage by about 9.4%.

Generally, the quantity $\%\Delta y/\Delta x$ is called the **semi-elasticity** of y with respect to x. The semi-elasticity is the percentage change in y when x increases by one *unit*. What we have just shown is that, in model (A.27), the semi-elasticity is constant and equal to $100\cdot\beta_1$. In Example A.6, we can conveniently summarize the relationship between wages and education by saying that one more year of education—starting from any amount of education—increases the wage by about 9.4%. This is why such models play an important role in economics.

Another relationship of some interest in applied economics is

$$y = \beta_0 + \beta_1\log(x), \qquad [A.29]$$

where $x > 0$. How can we interpret this equation? If we take the change in y, we get $\Delta y = \beta_1\Delta\log(x)$, which can be rewritten as $\Delta y = (\beta_1/100)[100\cdot\Delta\log(x)]$. Thus, using the approximation in (A.23), we have

$$\Delta y \approx (\beta_1/100)(\%\Delta x). \qquad [A.30]$$

In other words, $\beta_1/100$ is the unit change in y when x increases by 1%.

EXAMPLE A.7 **Labor Supply Function**

Assume that the labor supply of a worker can be described by

$$hours = 33 + 45.1\log(wage),$$

where *wage* is hourly wage and *hours* is hours worked per week. Then, from (A.30),

$$\Delta hours \approx (45.1/100)(\%\Delta wage) = .451\ \%\Delta wage.$$

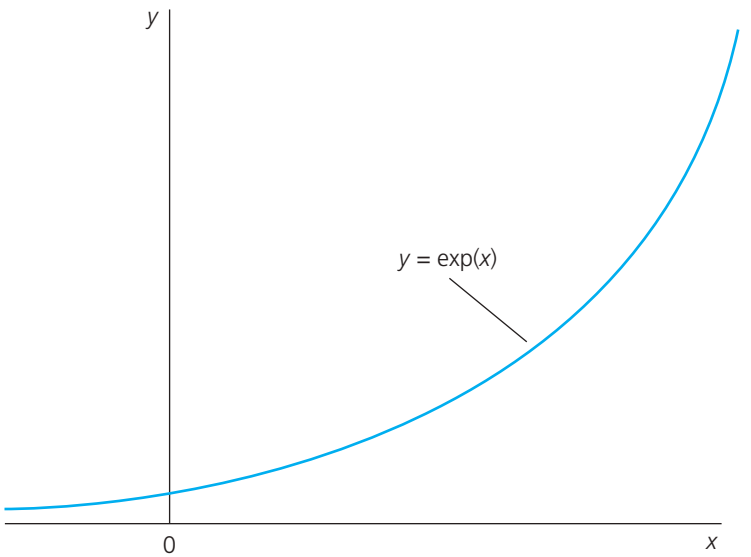
In other words, a 1% increase in *wage* increases the weekly hours worked by about .45, or slightly less than one-half hour. If the wage increases by 10%, then $\Delta hours = .451(10) = 4.51$, or about four and one-half hours. We would not want to use this approximation for much larger percentage changes in wages.

A-4c The Exponential Function

Before leaving this section, we need to discuss a special function that is related to the log. As motivation, consider equation (A.27). There, $\log(y)$ is a linear function of x. But how do we find y itself as a function of x? The answer is given by the **exponential function**.

We will write the exponential function as $y = \exp(x)$, which is graphed in Figure A.5. From Figure A.5, we see that $\exp(x)$ is defined for any value of x and is always greater than zero. Sometimes,

FIGURE A.5 Graph of $y = \exp(x)$.



the exponential function is written as $y = e^x$, but we will not use this notation. Two important values of the exponential function are $\exp(0) = 1$ and $\exp(1) = 2.7183$ (to four decimal places).

The exponential function is the inverse of the log function in the following sense: $\log[\exp(x)] = x$ for all x, and $\exp[\log(x)] = x$ for $x > 0$. In other words, the log "undoes" the exponential, and vice versa. (This is why the exponential function is sometimes called the *anti-log* function.) In particular, note that $\log(y) = \beta_0 + \beta_1 x$ is equivalent to

$$y = \exp(\beta_0 + \beta_1 x).$$

If $\beta_1 > 0$, the relationship between x and y has the same shape as in Figure A.5. Thus, if $\log(y) = \beta_0 + \beta_1 x$ with $\beta_1 > 0$, then x has an *increasing* marginal effect on y. In Example A.6, this means that another year of education leads to a larger change in wage than the previous year of education.

Two useful facts about the exponential function are $\exp(x_1 + x_2) = \exp(x_1)\exp(x_2)$ and $\exp[c \cdot \log(x)] = x^c$.

A-5 Differential Calculus

In the previous section, we asserted several approximations that have foundations in calculus. Let $y = f(x)$ for some function f. Then, for small changes in x,

$$\Delta y \approx \frac{df}{dx} \cdot \Delta x, \qquad \text{[A.31]}$$

where df/dx is the derivative of the function f, evaluated at the initial point x_0. We also write the derivative as dy/dx.

For example, if $y = \log(x)$, then $dy/dx = 1/x$. Using (A.31), with dy/dx evaluated at x_0, we have $\Delta y \approx (1/x_0)\Delta x$, or $\Delta\log(x) \approx \Delta x/x_0$, which is the approximation given in (A.22).

In applying econometrics, it helps to recall the derivatives of a handful of functions because we use the derivative to define the slope of a function at a given point. We can then use (A.31) to find the approximate change in y for small changes in x. In the linear case, the derivative is simply the slope of the line, as we would hope: if $y = \beta_0 + \beta_1 x$, then $dy/dx = \beta_1$.

If $y = x^c$, then $dy/dx = cx^{c-1}$. The derivative of a sum of two functions is the sum of the derivatives: $d[f(x) + g(x)]/dx = df(x)/dx + dg(x)/dx$. The derivative of a constant times any function is that same constant times the derivative of the function: $d[cf(x)]/dx = c[df(x)/dx]$. These simple rules allow us to find derivatives of more complicated functions. Other rules, such as the product, quotient, and chain rules, will be familiar to those who have taken calculus, but we will not review those here.

Some functions that are often used in economics, along with their derivatives, are

$$y = \beta_0 + \beta_1 x + \beta_2 x^2;\ dy/dx = \beta_1 + 2\beta_2 x$$
$$y = \beta_0 + \beta_1/x;\ dy/dx = -\beta_1/(x^2)$$
$$y = \beta_0 + \beta_1\sqrt{x};\ dy/dx = (\beta_1/2)x^{-1/2}$$
$$y = \beta_0 + \beta_1\log(x);\ dy/dx = \beta_1/x$$
$$y = \exp(\beta_0 + \beta_1 x);\ dy/dx = \beta_1\exp(\beta_0 + \beta_1 x).$$

If $\beta_0 = 0$ and $\beta_1 = 1$ in this last expression, we get $dy/dx = \exp(x)$, when $y = \exp(x)$.

In Section A-4, we noted that equation (A.26) defines a constant elasticity model when calculus is used. The calculus definition of elasticity is $(dy/dx)\cdot(x/y)$. It can be shown using properties of logs and exponentials that, when (A.26) holds, $(dy/dx)\cdot(x/y) = \beta_1$.

When y is a function of multiple variables, the notion of a **partial derivative** becomes important. Suppose that

$$y = f(x_1, x_2). \qquad \text{[A.32]}$$

Then, there are two partial derivatives, one with respect to x_1 and one with respect to x_2. The partial derivative of y with respect to x_1, denoted here by $\partial y/\partial x_1$, is just the usual derivative of (A.32) with respect to x_1, where x_2 is treated as a *constant*. Similarly, $\partial y/\partial x_2$ is just the derivative of (A.32) with respect to x_2, holding x_1 fixed.

Partial derivatives are useful for much the same reason as ordinary derivatives. We can approximate the change in y as

$$\Delta y \approx \frac{\partial y}{\partial x_1}\cdot\Delta x_1, \text{ holding } x_2 \text{ fixed.} \qquad \text{[A.33]}$$

Thus, calculus allows us to define partial effects in nonlinear models just as we could in linear models. In fact, if

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

then

$$\frac{\partial y}{\partial x_1} = \beta_1, \frac{\partial y}{\partial x_2} = \beta_2.$$

These can be recognized as the partial effects defined in Section A-2.

A more complicated example is

$$y = 5 + 4x_1 + x_1^2 - 3x_2 + 7x_1\cdot x_2. \qquad \text{[A.34]}$$

Now, the derivative of (A.34), with respect to x_1 (treating x_2 as a constant), is simply

$$\frac{\partial y}{\partial x_1} = 4 + 2x_1 + 7x_2;$$

note how this depends on x_1 and x_2. The derivative of (A.34), with respect to x_2, is $\partial y/\partial x_2 = -3 + 7x_1$, so this depends only on x_1.

EXAMPLE A.8 Wage Function with Interaction

A function relating wages to years of education and experience is

$$wage = 3.10 + .41\ educ + .19\ exper - .004\ exper^2 + .007\ educ \cdot exper. \qquad [A.35]$$

The partial effect of *exper* on *wage* is the partial derivative of (A.35):

$$\frac{\partial wage}{\partial exper} = .19 - .008\ exper + .007\ educ.$$

This is the approximate change in wage due to increasing experience by one year. Notice that this partial effect depends on the initial level of *exper* and *educ*. For example, for a worker who is starting with $educ = 12$ and $exper = 5$, the next year of experience increases wage by about $.19 - .008(5) + .007(12) = .234$, or 23.4 cents per hour. The exact change can be calculated by computing (A.35) at $exper = 5$, $educ = 12$ and at $exper = 6$, $educ = 12$, and then taking the difference. This turns out to be .23, which is very close to the approximation.

Differential calculus plays an important role in minimizing and maximizing functions of one or more variables. If $f(x_1, x_2, \ldots, x_k)$ is a differentiable function of k variables, then a necessary condition for $x_1^*, x_2^*, \ldots, x_k^*$ to either minimize or maximize f over all possible values of x_j is

$$\frac{\partial f}{\partial x_i}(x_1^*, x_2^*, \ldots, x_k^*) = 0, j = 1, 2, \ldots, k. \qquad [A.36]$$

In other words, all of the partial derivatives of f must be zero when they are evaluated at the x_h*. These are called the *first order conditions* for minimizing or maximizing a function. Practically, we hope to solve equation (A.36) for the x_h*. Then, we can use other criteria to determine whether we have minimized or maximized the function. We will not need those here. [See Sydsaeter and Hammond (1995) for a discussion of multivariable calculus and its use in optimizing functions.]

Summary

The math tools reviewed here are crucial for understanding regression analysis and the probability and statistics that are covered in Appendices B and C. The material on nonlinear functions—especially quadratic, logarithmic, and exponential functions—is critical for understanding modern applied economic research. The level of comprehension required of these functions does not include a deep knowledge of calculus, although calculus is needed for certain derivations.

Key Terms

Average
Ceteris Paribus
Constant Elasticity Model
Derivative
Descriptive Statistic
Diminishing Marginal Effect
Elasticity
Exponential Function
Intercept
Linear Function
Log Function
Marginal Effect
Median
Natural Logarithm
Nonlinear Function
Partial Derivative
Partial Effect
Percentage Change
Percentage Point Change
Proportionate Change
Relative Change
Semi-Elasticity
Slope
Summation Operator

Problems

1 The following table contains monthly housing expenditures for 10 families.

Family	Monthly Housing Expenditures (Dollars)
1	300
2	440
3	350
4	1,100
5	640
6	480
7	450
8	700
9	670
10	530

(i) Find the average monthly housing expenditure.
(ii) Find the median monthly housing expenditure.
(iii) If monthly housing expenditures were measured in hundreds of dollars, rather than in dollars, what would be the average and median expenditures?
(iv) Suppose that family number 8 increases its monthly housing expenditure to $900, but the expenditures of all other families remain the same. Compute the average and median housing expenditures.

2 Suppose the following equation describes the relationship between the average number of classes missed during a semester (*missed*) and the distance from school (*distance*, measured in miles):

$$missed = 3 + 0.2\, distance.$$

(i) Sketch this line, being sure to label the axes. How do you interpret the intercept in this equation?
(ii) What is the average number of classes missed for someone who lives five miles away?
(iii) What is the difference in the average number of classes missed for someone who lives 10 miles away and someone who lives 20 miles away?

3 In Example A.2, quantity of compact discs was related to price and income by $quantity = 120 - 9.8\,price + .03\,income$. What is the demand for CDs if $price = 15$ and $income = 200$? What does this suggest about using linear functions to describe demand curves?

4 Suppose the unemployment rate in the United States goes from 6.4% in one year to 5.6% in the next.
(i) What is the percentage point decrease in the unemployment rate?
(ii) By what percentage has the unemployment rate fallen?

5 Suppose that the return from holding a particular firm's stock goes from 15% in one year to 18% in the following year. The majority shareholder claims that "the stock return only increased by 3%," while the chief executive officer claims that "the return on the firm's stock increased by 20%." Reconcile their disagreement.

6 Suppose that Person A earns $35,000 per year and Person B earns $42,000.
(i) Find the exact percentage by which Person B's salary exceeds Person A's.
(ii) Now, use the difference in natural logs to find the approximate percentage difference.

7 Suppose the following model describes the relationship between annual salary (*salary*) and the number of previous years of labor market experience (*exper*):

$$\log(salary) = 10.6 + .027\,exper.$$

(i) What is *salary* when $exper = 0$? When $exper = 5$? (*Hint*: You will need to exponentiate.)
(ii) Use equation (A.28) to approximate the percentage increase in *salary* when *exper* increases by five years.
(iii) Use the results of part (i) to compute the exact percentage difference in salary when $exper = 5$ and $exper = 0$. Comment on how this compares with the approximation in part (ii).

8 Let *grthemp* denote the proportionate growth in employment, at the county level, from 1990 to 1995, and let *salestax* denote the county sales tax rate, stated as a proportion. Interpret the intercept and slope in the equation

$$grthemp = .043 - .78\,salestax.$$

9 Suppose the yield of a certain crop (in bushels per acre) is related to fertilizer amount (in pounds per acre) as

$$yield = 120 + .19\sqrt{fertilizer}.$$

(i) Graph this relationship by plugging in several values for *fertilizer*.
(ii) Describe how the shape of this relationship compares with a linear relationship between *yield* and *fertilizer*.

10 Suppose that in a particular state a standardized test is given to all graduating seniors. Let *score* denote a student's score on the test. Someone discovers that performance on the test is related to the size of the student's graduating high school class. The relationship is quadratic:

$$score = 45.6 + .082\,class - .000147\,class^2,$$

where *class* is the number of students in the graduating class.
(i) How do you literally interpret the value 45.6 in the equation? By itself, is it of much interest? Explain.
(ii) From the equation, what is the optimal size of the graduating class (the size that maximizes the test score)? (Round your answer to the nearest integer.) What is the highest achievable test score?

(iii) Sketch a graph that illustrates your solution in part (ii).
(iv) Does it seem likely that *score* and *class* would have a deterministic relationship? That is, is it realistic to think that once you know the size of a student's graduating class you know, with certainty, his or her test score? Explain.

11 Consider the line

$$y = \beta_0 + \beta_1 x.$$

(i) Let (x_1, y_1) and (x_2, y_2) be two points on the line. Show that $(\bar{x}, \bar{y})$ is also on the line, where $\bar{x} = (x_1 + x_2)/2$ is the average of the two values and $\bar{y} = (y_1 + y_2)/2$.
(ii) Extend the result of part (i) to n points on the line, $\{(x_i, y_i): i = 1, \ldots, n\}$.

12 (i) Let $\{x_i: i = 1, 2, \ldots, n\}$ be a set of n data points, and let $\bar{x}$ be the average. Suppose that the units i are divided into two groups of sizes n_1 and n_2, with $n_1 + n_2 = n$. Without loss of generality, order the observations as

$$\{x_1, x_2, \ldots, x_{n1}, x_{n1+1}, x_{n1+2}, \ldots, x_n\},$$

so that the data points for the first group appear first. Let

$$\bar{x}_1 = n_1^{-1} \sum_{i=1}^{n_1} x_i, \quad \bar{x}_2 = n_2^{-1} \sum_{i=n_1+1}^{n} x_i$$

be the averages for the two groups. Show that

$$\bar{x} = \left(\frac{n_1}{n}\right)\bar{x}_1 + \left(\frac{n_2}{n}\right)\bar{x}_2 = w_1\bar{x}_1 + w_2\bar{x}_2,$$

so that $\bar{x}$ can be expressed as a weighted average of the averages from the two subgroups.
(ii) Do the weights w_1 and w_2 in part (i) make intuitive sense? Explain.
(iii) How does the finding in part (i) extend the case of g groups, where the group sizes are $n_1, n_2, \ldots, n_g$?

13 (i) Let $\{x_i: i = 1, 2, \ldots, n\}$ be a set of n data points with $x_i > 0$ for all i. Is it always true that

$$\sum_{i=1}^{n} \frac{1}{x_i} = \frac{1}{\sum_{i=1}^{n} x_i}?$$

(ii) Is the equality in part (i) always true if $x_i = c$ for all i,where $c > 0$?

Math Refresher B

Fundamentals of Probability

This Math Refresher covers key concepts from basic probability. Appendices B and C are primarily for review; they are not intended to replace a course in probability and statistics. However, all of the probability and statistics concepts that we use in the text are covered in these appendices.

Probability is of interest in its own right for students in business, economics, and other social sciences. For example, consider the problem of an airline trying to decide how many reservations to accept for a flight that has 100 available seats. If fewer than 100 people want reservations, then these should all be accepted. But what if more than 100 people request reservations? A safe solution is to accept at most 100 reservations. However, because some people book reservations and then do not show up for the flight, there is some chance that the plane will not be full even if 100 reservations are booked. This results in lost revenue to the airline. A different strategy is to book more than 100 reservations and to hope that some people do not show up, so the final number of passengers is as close to 100 as possible. This policy runs the risk of the airline having to compensate people who are necessarily bumped from an overbooked flight.

A natural question in this context is: Can we decide on the optimal (or best) number of reservations the airline should make? This is a nontrivial problem. Nevertheless, given certain information (on airline costs and how frequently people show up for reservations), we can use basic probability to arrive at a solution.

B-1 Random Variables and Their Probability Distributions

Suppose that we flip a coin 10 times and count the number of times the coin turns up heads. This is an example of an **experiment**. Generally, an experiment is any procedure that can, at least in theory, be infinitely repeated and has a well-defined set of outcomes. We could, in principle, carry out the coin-flipping procedure again and again. Before we flip the coin, we know that the number of heads appearing is an integer from 0 to 10, so the outcomes of the experiment are well defined.

A **random variable** is one that takes on numerical values and has an outcome that is determined by an experiment. In the coin-flipping example, the number of heads appearing in 10 flips of a coin is an example of a random variable. Before we flip the coin 10 times, we do not know how many

times the coin will come up heads. Once we flip the coin 10 times and count the number of heads, we obtain the outcome of the random variable for this particular trial of the experiment. Another trial can produce a different outcome.

In the airline reservation example mentioned earlier, the number of people showing up for their flight is a random variable: before any particular flight, we do not know how many people will show up.

To analyze data collected in business and the social sciences, it is important to have a basic understanding of random variables and their properties. Following the usual conventions in probability and statistics throughout Appendices B and C, we denote random variables by uppercase letters, usually W, X, Y, and Z; particular outcomes of random variables are denoted by the corresponding lowercase letters, w, x, y, and z. For example, in the coin-flipping experiment, let X denote the number of heads appearing in 10 flips of a coin. Then, X is not associated with any particular value, but we know X will take on a value in the set $\{0, 1, 2, \ldots, 10\}$. A particular outcome is, say, $x = 6$.

We indicate large collections of random variables by using subscripts. For example, if we record last year's income of 20 randomly chosen households in the United States, we might denote these random variables by $X_1, X_2, \ldots, X_{20}$; the particular outcomes would be denoted $x_1, x_2, \ldots, x_{20}$.

As stated in the definition, random variables are always defined to take on numerical values, even when they describe qualitative events. For example, consider tossing a single coin, where the two outcomes are heads and tails. We can define a random variable as follows: $X = 1$ if the coin turns up heads, and $X = 0$ if the coin turns up tails.

A random variable that can only take on the values zero and one is called a **Bernoulli** (or **binary**) **random variable**. In basic probability, it is traditional to call the event $X = 1$ a "success" and the event $X = 0$ a "failure." For a particular application, the success-failure nomenclature might not correspond to our notion of a success or failure, but it is a useful terminology that we will adopt.

B-1a Discrete Random Variables

A **discrete random variable** is one that takes on only a finite or countably infinite number of values. The notion of "countably infinite" means that even though an infinite number of values can be taken on by a random variable, those values can be put in a one-to-one correspondence with the positive integers. Because the distinction between "countably infinite" and "uncountably infinite" is somewhat subtle, we will concentrate on discrete random variables that take on only a finite number of values. Larsen and Marx (1986, Chapter 3) provide a detailed treatment.

A Bernoulli random variable is the simplest example of a discrete random variable. The only thing we need to completely describe the behavior of a Bernoulli random variable is the probability that it takes on the value one. In the coin-flipping example, if the coin is "fair," then $P(X = 1) = 1/2$ (read as "the probability that X equals one is one-half"). Because probabilities must sum to one, $P(X = 0) = 1/2$, also.

Social scientists are interested in more than flipping coins, so we must allow for more general situations. Again, consider the example where the airline must decide how many people to book for a flight with 100 available seats. This problem can be analyzed in the context of several Bernoulli random variables as follows: for a randomly selected customer, define a Bernoulli random variable as $X = 1$ if the person shows up for the reservation, and $X = 0$ if not.

There is no reason to think that the probability of any particular customer showing up is 1/2; in principle, the probability can be any number between 0 and 1. Call this number θ, so that

$$P(X = 1) = \theta \qquad \text{[B.1]}$$

$$P(X = 0) = 1 - \theta. \qquad \text{[B.2]}$$

For example, if $\theta = .75$, then there is a 75% chance that a customer shows up after making a reservation and a 25% chance that the customer does not show up. Intuitively, the value of θ is crucial in determining the airline's strategy for booking reservations. Methods for *estimating* θ, given historical data on airline reservations, are a subject of mathematical statistics, something we turn to in Math Refresher C.

More generally, any discrete random variable is completely described by listing its possible values and the associated probability that it takes on each value. If X takes on the k possible values $\{x_1, \ldots, x_k\}$, then the probabilities $p_1, p_2, \ldots, p_k$ are defined by

$$p_j = \mathrm{P}(X = x_j), j = 1, 2, \ldots, k, \qquad \text{[B.3]}$$

where each p_j is between 0 and 1 and

$$p_1 + p_2 + \cdots + p_k = 1. \qquad \text{[B.4]}$$

Equation (B.3) is read as: "The probability that X takes on the value x_j is equal to p_j."

Equations (B.1) and (B.2) show that the probabilities of success and failure for a Bernoulli random variable are determined entirely by the value of θ. Because Bernoulli random variables are so prevalent, we have a special notation for them: $X \sim \text{Bernoulli}(\theta)$ is read as "X has a Bernoulli distribution with probability of success equal to θ."

The **probability density function (pdf)** of X summarizes the information concerning the possible outcomes of X and the corresponding probabilities:

$$f(x_j) = p_j, j = 1, 2, \ldots, k, \qquad \text{[B.5]}$$

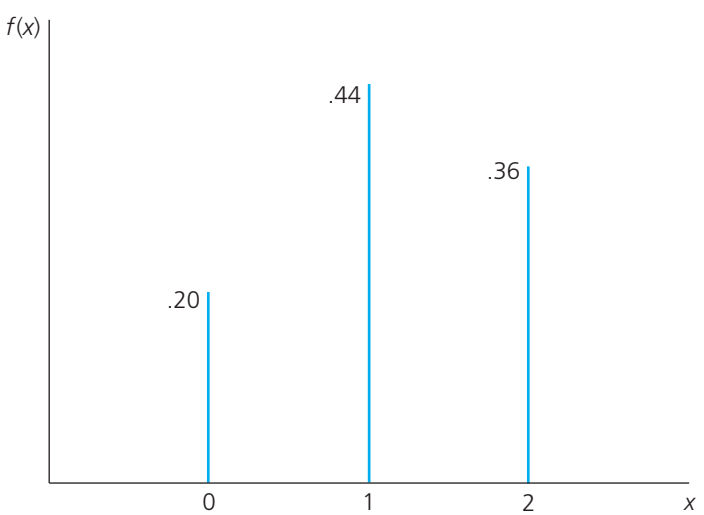
with $f(x) = 0$ for any x not equal to x_j for some j. In other words, for any real number x, $f(x)$ is the probability that the random variable X takes on the particular value x. When dealing with more than one random variable, it is sometimes useful to subscript the pdf in question: f_X is the pdf of X, f_Y is the pdf of Y, and so on.

Given the pdf of any discrete random variable, it is simple to compute the probability of any event involving that random variable. For example, suppose that X is the number of free throws made by a basketball player out of two attempts, so that X can take on the three values $\{0, 1, 2\}$. Assume that the pdf of X is given by

$$f(0) = .20, f(1) = .44, \text{ and } f(2) = .36.$$

The three probabilities sum to one, as they must. Using this pdf, we can calculate the probability that the player makes *at least* one free throw: $\mathrm{P}(X \geq 1) = \mathrm{P}(X = 1) + \mathrm{P}(X = 2) = .44 + .36 = .80$. The pdf of X is shown in Figure B.1.

FIGURE B.1 The pdf of the number of free throws made out of two attempts.



B-1b Continuous Random Variables

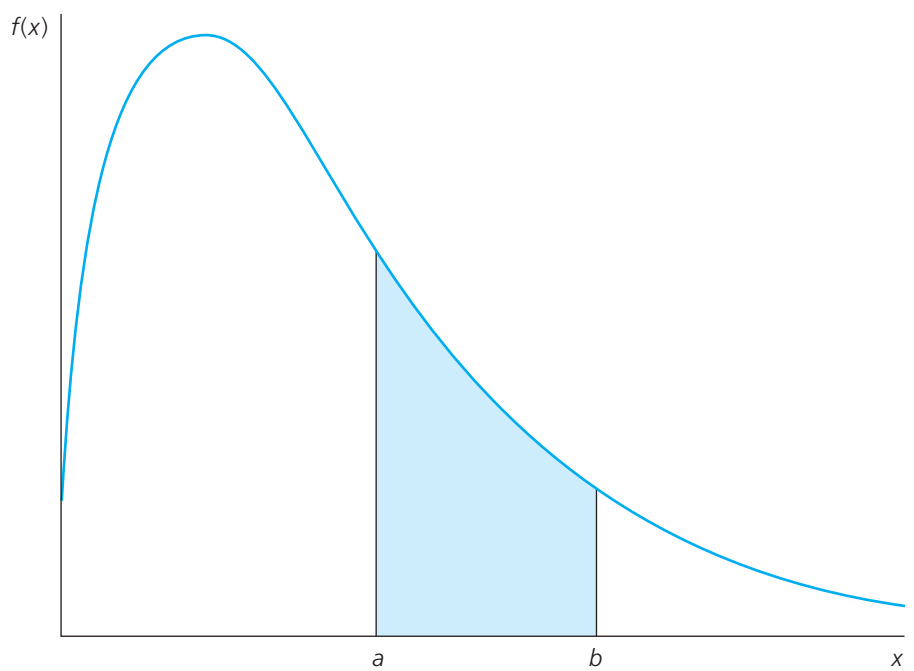
A variable X is a **continuous random variable** if it takes on any real value with *zero* probability. This definition is somewhat counterintuitive because in any application we eventually observe some outcome for a random variable. The idea is that a continuous random variable X can take on so many possible values that we cannot count them or match them up with the positive integers, so logical consistency dictates that X can take on each value with probability zero. While measurements are always discrete in practice, random variables that take on numerous values are best treated as continuous. For example, the most refined measure of the price of a good is in terms of cents. We can imagine listing all possible values of price in order (even though the list may continue indefinitely), which technically makes price a discrete random variable. However, there are so many possible values of price that using the mechanics of discrete random variables is not feasible.

We can define a probability density function for continuous random variables, and, as with discrete random variables, the pdf provides information on the likely outcomes of the random variable. However, because it makes no sense to discuss the probability that a continuous random variable takes on a particular value, we use the pdf of a continuous random variable only to compute events involving a range of values. For example, if a and b are constants where $a < b$, the probability that X lies between the numbers a and b, $\mathrm{P}(a \leq X \leq b)$, is the *area* under the pdf between points a and b, as shown in Figure B.2. If you are familiar with calculus, you recognize this as the *integral* of the function f between the points a and b. The entire area under the pdf must always equal one.

When computing probabilities for continuous random variables, it is easiest to work with the **cumulative distribution function (cdf)**. If X is any random variable, then its cdf is defined for any real number x by

$$F(x) \equiv \mathrm{P}(X \leq x). \qquad \text{[B.6]}$$

FIGURE B.2 The probability that *X* lies between the points *a* and *b*.



For discrete random variables, (B.6) is obtained by summing the pdf over all values x_j such that $x_j \leq x$. For a continuous random variable, $F(x)$ is the area under the pdf, f, to the left of the point x. Because $F(x)$ is simply a probability, it is always between 0 and 1. Further, if $x_1 < x_2$, then $P(X \leq x_1) \leq P(X \leq x_2)$, that is, $F(x_1) \leq F(x_2)$. This means that a cdf is an increasing (or at least a nondecreasing) function of x.

Two important properties of cdfs that are useful for computing probabilities are the following:

$$\text{For any number } c,\ P(X > c) = 1 - F(c). \qquad \text{[B.7]}$$

$$\text{For any numbers } a < b,\ P(a < X \leq b) = F(b) - F(a). \qquad \text{[B.8]}$$

In our study of econometrics, we will use cdfs to compute probabilities only for continuous random variables, in which case it does not matter whether inequalities in probability statements are strict or not. That is, for a continuous random variable X,

$$P(X \geq c) = P(X > c), \qquad \text{[B.9]}$$

and

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b). \qquad \text{[B.10]}$$

Combined with (B.7) and (B.8), equations (B.9) and (B.10) greatly expand the probability calculations that can be done using continuous cdfs.

Cumulative distribution functions have been tabulated for all of the important continuous distributions in probability and statistics. The most well known of these is the normal distribution, which we cover along with some related distributions in Section B-5.

B-2 Joint Distributions, Conditional Distributions, and Independence

In economics, we are usually interested in the occurrence of events involving more than one random variable. For example, in the airline reservation example referred to earlier, the airline might be interested in the probability that a person who makes a reservation shows up *and* is a business traveler; this is an example of a *joint probability*. Or, the airline might be interested in the following *conditional probability*: conditional on the person being a business traveler, what is the probability of his or her showing up? In the next two subsections, we formalize the notions of joint and conditional distributions and the important notion of *independence* of random variables.

B-2a Joint Distributions and Independence

Let X and Y be discrete random variables. Then, (X, Y) have a **joint distribution**, which is fully described by the *joint probability density function* of (X, Y):

$$f_{X,Y}(x, y) = P(X = x, Y = y), \qquad \text{[B.11]}$$

where the right-hand side is the probability that $X = x$ and $Y = y$. When X and Y are continuous, a joint pdf can also be defined, but we will not cover such details because joint pdfs for continuous random variables are not used explicitly in this text.

In one case, it is easy to obtain the joint pdf if we are given the pdfs of X and Y. In particular, random variables X and Y are said to be independent if, and only if,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \qquad \text{[B.12]}$$

for all x and y, where f_X is the pdf of X and f_Y is the pdf of Y. In the context of more than one random variable, the pdfs f_X and f_Y are often called *marginal probability density functions* to distinguish them from the joint pdf $f_{X,Y}$. This definition of independence is valid for discrete and continuous random variables.

To understand the meaning of (B.12), it is easiest to deal with the discrete case. If X and Y are discrete, then (B.12) is the same as

$$P(X = x, Y = y) = P(X = x)P(Y = y); \qquad [B.13]$$

in other words, the probability that $X = x$ and $Y = y$ is the product of the two probabilities $P(X = x)$ and $P(Y = y)$. One implication of (B.13) is that joint probabilities are fairly easy to compute, because they only require knowledge of $P(X = x)$ and $P(Y = y)$.

If random variables are not independent, then they are said to be *dependent*.

EXAMPLE B.1 **Free Throw Shooting**

Consider a basketball player shooting two free throws. Let X be the Bernoulli random variable equal to one if she or he makes the first free throw, and zero otherwise. Let Y be a Bernoulli random variable equal to one if he or she makes the second free throw. Suppose that she or he is an 80% free throw shooter, so that $P(X = 1) = P(Y = 1) = .8$. What is the probability of the player making both free throws?

If X and Y are independent, we can easily answer this question: $P(X = 1, Y = 1) = P(X = 1)P(Y = 1) = (.8)(.8) = .64$. Thus, there is a 64% chance of making both free throws. If the chance of making the second free throw depends on whether the first was made—that is, X and Y are not independent—then this simple calculation is not valid.

Independence of random variables is a very important concept. In the next subsection, we will show that if X and Y are independent, then knowing the outcome of X does not change the probabilities of the possible outcomes of Y, and vice versa. One useful fact about independence is that if X and Y are independent and we define new random variables $g(X)$ and $h(Y)$ for any functions g and h, then these new random variables are also independent.

There is no need to stop at two random variables. If $X_1, X_2, \ldots, X_n$ are discrete random variables, then their joint pdf is $f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. The random variables $X_1, X_2, \ldots, X_n$ are **independent random variables** if, and only if, their joint pdf is the product of the individual pdfs for any $(x_1, x_2, \ldots, x_n)$. This definition of independence also holds for continuous random variables.

The notion of independence plays an important role in obtaining some of the classic distributions in probability and statistics. Earlier, we defined a Bernoulli random variable as a zero-one random variable indicating whether or not some event occurs. Often, we are interested in the number of successes in a sequence of *independent* Bernoulli trials. A standard example of independent Bernoulli trials is flipping a coin again and again. Because the outcome on any particular flip has nothing to do with the outcomes on other flips, independence is an appropriate assumption.

Independence is often a reasonable approximation in more complicated situations. In the airline reservation example, suppose that the airline accepts n reservations for a particular flight. For each $i = 1, 2, \ldots, n$, let Y_i denote the Bernoulli random variable indicating whether customer i shows up: $Y_i = 1$ if customer i appears, and $Y_i = 0$ otherwise. Letting θ again denote the probability of success (using reservation), each Y_i has a Bernoulli(θ) distribution. As an approximation, we might assume that the Y_i are independent of one another, although this is not exactly true in reality: some people travel in groups, which means that whether or not a person shows up is not truly independent of whether all others show up. Modeling this kind of dependence is complex, however, so we might be willing to use independence as an approximation.

The variable of primary interest is the total number of customers showing up out of the n reservations; call this variable X. Because each Y_i is unity when a person shows up, we can write

$X = Y_1 + Y_2 + \cdots + Y_n$. Now, assuming that each Y_i has probability of success θ and that the Y_i are independent, X can be shown to have a **binomial distribution**. That is, the probability density function of X is

$$f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x = 0, 1, 2, \ldots, n, \qquad \text{[B.14]}$$

where $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$, and for any integer n, $n!$ (read "n factorial") is defined as $n! = n \cdot (n-1) \cdot (n-2) \cdots 1$. By convention, $0! = 1$. When a random variable X has the pdf given in (B.14), we write $X \sim \text{Binomial}(n, \theta)$. Equation (B.14) can be used to compute $P(X = x)$ for any value of x from 0 to n.

If the flight has 100 available seats, the airline is interested in $P(X > 100)$. Suppose, initially, that $n = 120$, so that the airline accepts 120 reservations, and the probability that each person shows up is $\theta = .85$. Then, $P(X > 100) = P(X = 101) + P(X = 102) + \cdots + P(X = 120)$, and each of the probabilities in the sum can be found from equation (B.14) with $n = 120$, $\theta = .85$, and the appropriate value of x (101 to 120). This is a difficult hand calculation, but many statistical packages have commands for computing this kind of probability. In this case, the probability that more than 100 people will show up is about .659, which is probably more risk of overbooking than the airline wants to tolerate. If, instead, the number of reservations is 110, the probability of more than 100 passengers showing up is only about .024.

B-2b Conditional Distributions

In econometrics, we are usually interested in how one random variable, call it Y, is related to one or more other variables. For now, suppose that there is only one variable whose effects we are interested in, call it X. The most we can know about how X affects Y is contained in the **conditional distribution** of Y given X. This information is summarized by the *conditional probability density function*, defined by

$$f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x) \qquad \text{[B.15]}$$

for all values of x such that $f_X(x) > 0$. The interpretation of (B.15) is most easily seen when X and Y are discrete. Then,

$$f_{Y|X}(y|x) = P(Y = y|X = x), \qquad \text{[B.16]}$$

where the right-hand side is read as "the probability that $Y = y$ given that $X = x$." When Y is continuous, $f_{Y|X}(y|x)$ is not interpretable directly as a probability, for the reasons discussed earlier, but conditional probabilities are found by computing areas under the conditional pdf.

An important feature of conditional distributions is that, if X and Y are independent random variables, knowledge of the value taken on by X tells us nothing about the probability that Y takes on various values (and vice versa). That is, $f_{Y|X}(y|x) = f_Y(y)$, and $f_{X|Y}(x|y) = f_X(x)$.

EXAMPLE B.2 **Free Throw Shooting**

Consider again the basketball-shooting example, where two free throws are to be attempted. Assume that the conditional density is

$$f_{Y|X}(1|1) = .85, f_{Y|X}(0|1) = .15$$
$$f_{Y|X}(1|0) = .70, f_{Y|X}(0|0) = .30.$$

This means that the probability of the player making the second free throw depends on whether the first free throw was made: if the first free throw is made, the chance of making the second is .85; if the

first free throw is missed, the chance of making the second is .70. This implies that X and Y are *not* independent; they are dependent.

We can still compute $P(X = 1, Y = 1)$ provided we know $P(X = 1)$. Assume that the probability of making the first free throw is .8, that is, $P(X = 1) = .8$. Then, from (B.15), we have

$$P(X = 1, Y = 1) = P(Y = 1|X = 1) \cdot P(X = 1) = (.85)(.8) = .68.$$

B-3 Features of Probability Distributions

For many purposes, we will be interested in only a few aspects of the distributions of random variables. The features of interest can be put into three categories: measures of central tendency, measures of variability or spread, and measures of association between two random variables. We cover the last of these in Section B-4.

B-3a A Measure of Central Tendency: The Expected Value

The expected value is one of the most important probabilistic concepts that we will encounter in our study of econometrics. If X is a random variable, the **expected value** (or expectation) of X, denoted $E(X)$ and sometimes μX or simply μ, is a weighted average of all possible values of X. The weights are determined by the probability density function. Sometimes, the expected value is called the *population mean*, especially when we want to emphasize that X represents some variable in a population.

The precise definition of expected value is simplest in the case that X is a discrete random variable taking on a finite number of values, say, $\{x_1, \ldots, x_k\}$. Let $f(x)$ denote the probability density function of X. The expected value of X is the weighted average

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_k f(x_k) \equiv \sum_{j=1}^{k} x_j f(x_j). \qquad [B.17]$$

This is easily computed given the values of the pdf at each possible outcome of X.

EXAMPLE B.3 Computing an Expected Value

Suppose that X takes on the values -1, 0, and 2 with probabilities 1/8, 1/2, and 3/8, respectively. Then,

$$E(X) = (-1)\cdot(1/8) + 0\cdot(1/2) + 2\cdot(3/8) = 5/8.$$

This example illustrates something curious about expected values: the expected value of X can be a number that is not even a possible outcome of X. We know that X takes on the values -1, 0, or 2, yet its expected value is 5/8. This makes the expected value deficient for summarizing the central tendency of certain discrete random variables, but calculations such as those just mentioned can be useful, as we will see later.

If X is a continuous random variable, then $E(X)$ is defined as an integral:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx, \qquad [B.18]$$

which we assume is well defined. This can still be interpreted as a weighted average. For the most common continuous distributions, $E(X)$ is a number that is a possible outcome of X. In this text, we will not need to compute expected values using integration, although we will draw on some well-known results from probability for expected values of special random variables.

Given a random variable X and a function $g(\cdot)$, we can create a new random variable $g(X)$. For example, if X is a random variable, then so is X^2 and $\log(X)$(if $X > 0$). The expected value of $g(X)$ is, again, simply a weighted average:

$$\mathrm{E}[g(X)] = \sum_{j=1}^{k} g(x_j) f_X(x_j) \qquad [B.19]$$

or, for a continuous random variable,

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx. \qquad [B.20]$$

EXAMPLE B.4 Expected Value of X^2

For the random variable in Example B.3, let $g(X) = X^2$. Then,

$$\mathrm{E}(X^2) = (-1)^2(1/8) + (0)^2(1/2) + (2)^2(3/8) = 13/8.$$

In Example B.3, we computed $\mathrm{E}(X) = 5/8$, so that $[\mathrm{E}(X)]^2 = 25/64$. This shows that $\mathrm{E}(X^2)$ is *not* the same as $[\mathrm{E}(X)]^2$. In fact, for a nonlinear function $g(X)$, $\mathrm{E}[g(X)] \neq g[\mathrm{E}(X)]$ (except in very special cases).

If X and Y are random variables, then $g(X, Y)$ is a random variable for any function g, and so we can define its expectation. When X and Y are both discrete, taking on values $\{x_1, x_2, \ldots, x_k\}$ and $\{y_1, y_2, \ldots, y_m\}$, respectively, the expected value is

$$\mathrm{E}[g(X, Y)] = \sum_{h=1}^{k} \sum_{j=1}^{m} g(x_h, y_j) f_{X,Y}(x_h, y_j),$$

where $f_{X,Y}$ is the joint pdf of (X, Y). The definition is more complicated for continuous random variables because it involves integration; we do not need it here. The extension to more than two random variables is straightforward.

B-3b Properties of Expected Values

In econometrics, we are not so concerned with computing expected values from various distributions; the major calculations have been done many times, and we will largely take these on faith. We will need to manipulate some expected values using a few simple rules. These are so important that we give them labels:

Property E.1: For any constant c, $\mathrm{E}(c) = c$.

Property E.2: For any constants a and b, $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.

One useful implication of E.2 is that, if $\mu = \mathrm{E}(X)$, and we define a new random variable as $Y = X - \mu$, then $\mathrm{E}(Y) = 0$; in E.2, take $a = 1$ and $b = -\mu$.

As an example of Property E.2, let X be the temperature measured in Celsius at noon on a particular day at a given location; suppose the expected temperature is $\mathrm{E}(X) = 25$. If Y is the temperature measured in Fahrenheit, then $Y = 32 + (9/5)X$. From Property E.2, the expected temperature in Fahrenheit is $\mathrm{E}(Y) = 32 + (9/5)\cdot\mathrm{E}(X) = 32 + (9/5)\cdot 25 = 77$.

Generally, it is easy to compute the expected value of a linear function of many random variables.

Property E.3: If $\{a_1, a_2, \ldots, a_n\}$ are constants and $\{X_1, X_2, \ldots, X_n\}$ are random variables, then

$$\mathrm{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1\mathrm{E}(X_1) + a_2\mathrm{E}(X_2) + \cdots + a_n\mathrm{E}(X_n).$$

Or, using summation notation,

$$\mathrm{E}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i\mathrm{E}(X_i). \qquad [B.21]$$

As a special case of this, we have (with each $a_i = 1$)

$$\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i), \qquad [B.22]$$

so that the expected value of the sum is the sum of expected values. This property is used often for derivations in mathematical statistics.

EXAMPLE B.5 **Finding Expected Revenue**

Let X_1, X_2, and X_3 be the numbers of small, medium, and large pizzas, respectively, sold during the day at a pizza parlor. These are random variables with expected values $\mathrm{E}(X_1) = 25$, $\mathrm{E}(X_2) = 57$, and $\mathrm{E}(X_3) = 40$. The prices of small, medium, and large pizzas are \$5.50, \$7.60, and \$9.15. Therefore, the expected revenue from pizza sales on a given day is

$$\begin{aligned}\mathrm{E}(5.50\,X_1 + 7.60\,X_2 + 9.15\,X_3) &= 5.50\,\mathrm{E}(X_1) + 7.60\,\mathrm{E}(X_2) + 9.15\,\mathrm{E}(X_3)\\ &= 5.50(25) + 7.60(57) + 9.15(40) = 936.70,\end{aligned}$$

that is, \$936.70. The actual revenue on any particular day will generally differ from this value, but this is the *expected* revenue.

We can also use Property E.3 to show that if $X \sim \text{Binomial}(n, \theta)$, then $\mathrm{E}(X) = n\theta$. That is, the expected number of successes in n Bernoulli trials is simply the number of trials times the probability of success on any particular trial. This is easily seen by writing X as $X = Y_1 + Y_2 + \cdots + Y_n$, where each $Y_i \sim \text{Bernoulli}(\theta)$. Then,

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(Y_i) = \sum_{i=1}^{n} \theta = n\theta.$$

We can apply this to the airline reservation example, where the airline makes $n = 120$ reservations, and the probability of showing up is $\theta = .85$. The *expected* number of people showing up is $120(.85) = 102$. Therefore, if there are 100 seats available, the expected number of people showing up is too large; this has some bearing on whether it is a good idea for the airline to make 120 reservations.

Actually, what the airline should do is define a profit function that accounts for the net revenue earned per seat sold and the cost per passenger bumped from the flight. This profit function is random because the actual number of people showing up is random. Let r be the net revenue from each passenger. (You can think of this as the price of the ticket for simplicity.) Let c be the compensation owed to any passenger bumped from the flight. Neither r nor c is random; these are assumed to be known to the airline. Let Y denote profits for the flight. Then, with 100 seats available,

$$\begin{aligned}Y &= rX \text{ if } X \leq 100\\ &= 100r - c(X - 100) \text{ if } X > 100.\end{aligned}$$

The first equation gives profit if no more than 100 people show up for the flight; the second equation is profit if more than 100 people show up. (In the latter case, the net revenue from ticket sales is $100r$, because all 100 seats are sold, and then $c(X - 100)$ is the cost of making more than 100 reservations.) Using the fact that X has a Binomial(n,.85) distribution, where n is the number of reservations made, expected profits, E(Y), can be found as a function of n (and r and c). Computing E(Y) directly would be quite difficult, but it can be found quickly using a computer. Once values for r and c are given, the value of n that maximizes expected profits can be found by searching over different values of n.

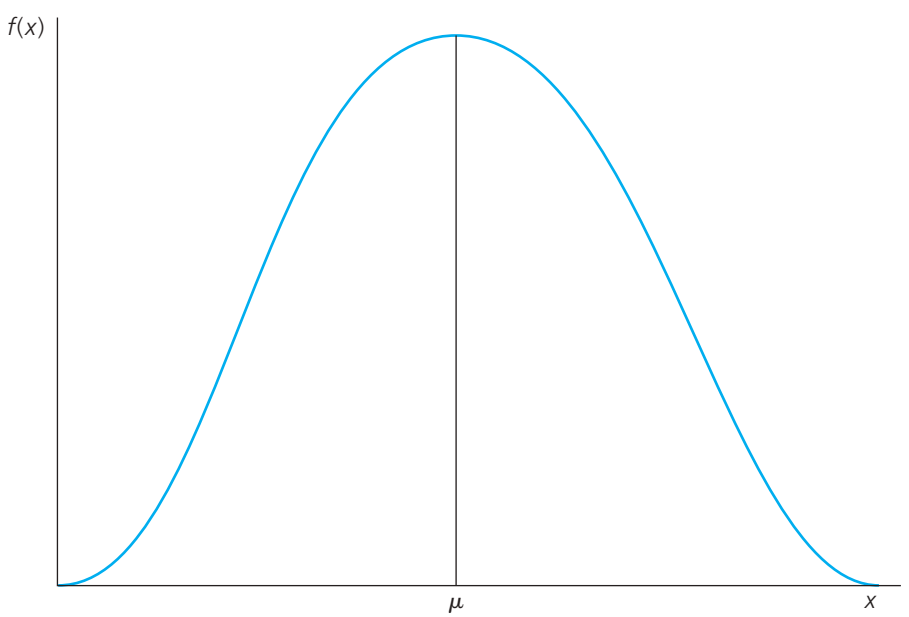
B-3c Another Measure of Central Tendency: The Median

The expected value is only one possibility for defining the central tendency of a random variable. Another measure of central tendency is the **median**. A general definition of *median* is too complicated for our purposes. If X is continuous, then the median of X, say, m, is the value such that one-half of the area under the pdf is to the left of m, and one-half of the area is to the right of m.

When X is discrete and takes on a finite number of odd values, the median is obtained by ordering the possible values of X and then selecting the value in the middle. For example, if X can take on the values $\{-4, 0, 2, 8, 10, 13, 17\}$, then the median value of X is 8. If X takes on an even number of values, there are really two median values; sometimes, these are averaged to get a unique median value. Thus, if X takes on the values $\{-5, 3, 9, 17\}$, then the median values are 3 and 9; if we average these, we get a median equal to 6.

In general, the median, sometimes denoted Med(X), and the expected value, E(X), are different. Neither is "better" than the other as a measure of central tendency; they are both valid ways to measure the center of the distribution of X. In one special case, the median and expected value (or mean) are the same. If X has a **symmetric distribution** about the value μ, then μ is both the expected value and the median. Mathematically, the condition is $f(\mu + x) = f(\mu - x)$ for all x. This case is illustrated in Figure B.3.

FIGURE B.3 A symmetric probability distribution.



B-3d Measures of Variability: Variance and Standard Deviation

Although the central tendency of a random variable is valuable, it does not tell us everything we want to know about the distribution of a random variable. Figure B.4 shows the pdfs of two random variables with the same mean. Clearly, the distribution of X is more tightly centered about its mean than is the distribution of Y. We would like to have a simple way of summarizing differences in the spreads of distributions.

B-3e Variance

For a random variable X, let $\mu = \mathrm{E}(X)$. There are various ways to measure how far X is from its expected value, but the simplest one to work with algebraically is the squared difference, $(X - \mu)^2$. (The squaring eliminates the sign from the distance measure; the resulting positive value corresponds to our intuitive notion of distance and treats values above and below μ symmetrically.) This distance is itself a random variable because it can change with every outcome of X. Just as we needed a number to summarize the central tendency of X, we need a number that tells us how far X is from μ, *on average*. One such number is the **variance**, which tells us the expected distance from X to its mean:

$$\mathrm{Var}(X) \equiv \mathrm{E}[(X - \mu)^2]. \qquad \text{[B.23]}$$

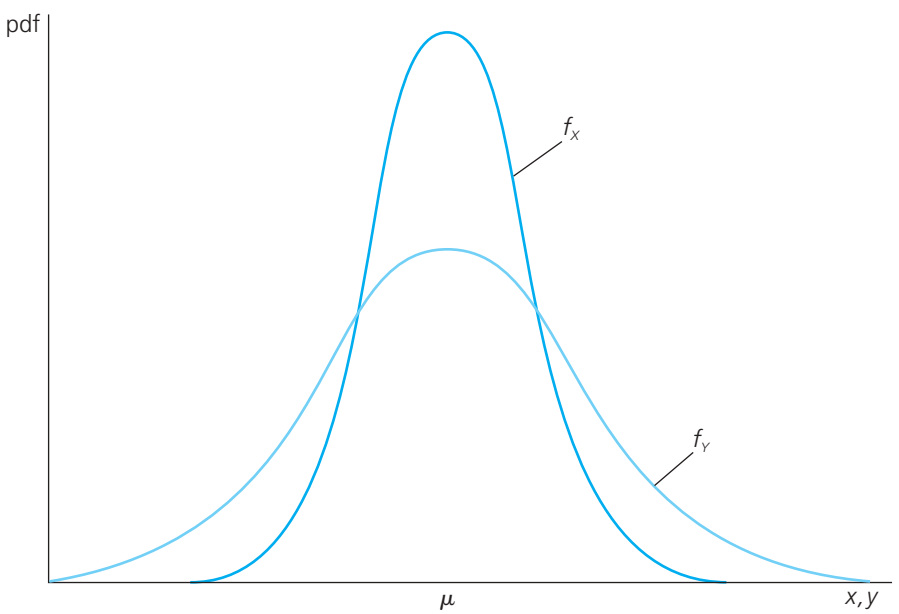
Variance is sometimes denoted σ_X^2, or simply σ^2, when the context is clear. From (B.23), it follows that the variance is always nonnegative.

As a computational device, it is useful to observe that

$$\sigma^2 = \mathrm{E}(X^2 - 2X\mu + \mu^2) = \mathrm{E}(X^2) - 2\mu^2 + \mu^2 = \mathrm{E}(X^2) - \mu^2. \qquad \text{[B.24]}$$

In using either (B.23) or (B.24), we need not distinguish between discrete and continuous random variables: the definition of variance is the same in either case. Most often, we first compute $\mathrm{E}(X)$, then $\mathrm{E}(X^2)$, and then we use the formula in (B.24). For example, if $X \sim \mathrm{Bernoulli}(\theta)$, then $\mathrm{E}(X) = \theta$, and, because $X^2 = X$, $\mathrm{E}(X^2) = \theta$. It follows from equation (B.24) that $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \theta - \theta^2 = \theta(1 - \theta)$.

FIGURE B.4 Random variables with the same mean but different distributions.



Two important properties of the variance follow.

Property VAR.1: $\text{Var}(X) = 0$ if, and only if, there is a constant c such that $P(X = c) = 1$, in which case $\text{E}(X) = c$.

This first property says that the variance of any constant is zero and if a random variable has zero variance, then it is essentially constant.

Property VAR.2: For any constants a and b, $\text{Var}(aX + b) = a^2\text{Var}(X)$.

This means that adding a constant to a random variable does not change the variance, but multiplying a random variable by a constant increases the variance by a factor equal to the *square* of that constant. For example, if X denotes temperature in Celsius and $Y = 32 + (9/5)X$ is temperature in Fahrenheit, then $\text{Var}(Y) = (9/5)^2\text{Var}(X) = (81/25)\text{Var}(X)$.

B-3f Standard Deviation

The **standard deviation** of a random variable, denoted $\text{sd}(X)$, is simply the positive square root of the variance: $\text{sd}(X) \equiv +\sqrt{\text{Var}(X)}$. The standard deviation is sometimes denoted σ_X, or simply σ, when the random variable is understood. Two standard deviation properties immediately follow from Properties VAR.1 and VAR.2.

Property SD.1: For any constant c, $\text{sd}(c) = 0$.

Property SD.2: For any constants a and b,

$$\text{sd}(aX + b) = |a|\text{sd}(X).$$

In particular, if $a > 0$, then $\text{sd}(aX) = a \cdot \text{sd}(X)$.

This last property makes the standard deviation more natural to work with than the variance. For example, suppose that X is a random variable measured in thousands of dollars, say, income. If we define $Y = 1{,}000X$, then Y is income measured in dollars. Suppose that $\text{E}(X) = 20$, and $\text{sd}(X) = 6$. Then, $\text{E}(Y) = 1{,}000\text{E}(X) = 20{,}000$, and $\text{sd}(Y) = 1{,}000 \cdot \text{sd}(X) = 6{,}000$, so that the expected value and standard deviation both increase by the same factor, 1,000. If we worked with variance, we would have $\text{Var}(Y) = (1{,}000)^2\text{Var}(X)$, so that the variance of Y is one million times larger than the variance of X.

B-3g Standardizing a Random Variable

As an application of the properties of variance and standard deviation—and a topic of practical interest in its own right—suppose that given a random variable X, we define a new random variable by subtracting off its mean m and dividing by its standard deviation σ:

$$Z \equiv \frac{X - \mu}{\sigma}, \qquad \text{[B.25]}$$

which we can write as $Z = aX + b$, where $a \equiv (1/\sigma)$ and $b \equiv -(\mu/\sigma)$. Then, from Property E.2,

$$\text{E}(Z) = a\text{E}(X) + b = (\mu/\sigma) - (\mu/\sigma) = 0.$$

From Property VAR.2,

$$\text{Var}(Z) = a^2\text{Var}(X) = (\sigma^2/\sigma^2) = 1.$$

Thus, the random variable Z has a mean of zero and a variance (and therefore a standard deviation) equal to one. This procedure is sometimes known as *standardizing* the random variable X, and Z is called a **standardized random variable**. (In introductory statistics courses, it is sometimes called the *z-transform* of X.) It is important to remember that the standard deviation, not the variance, appears in the denominator of (B.25). As we will see, this transformation is frequently used in statistical inference.

As a specific example, suppose that $E(X) = 2$, and $Var(X) = 9$. Then, $Z = (X - 2)/3$ has expected value zero and variance one.

B-3h Skewness and Kurtosis

We can use the standardized version of a random variable to define other features of the distribution of a random variable. These features are described by using what are called *higher order moments*. For example, the third moment of the random variable Z in (B.25) is used to determine whether a distribution is symmetric about its mean. We can write

$$E(Z^3) = E[(X - \mu)^3]/\sigma^3.$$

If X has a symmetric distribution about μ, then Z has a symmetric distribution about zero. (The division by σ^3 does not change whether the distribution is symmetric.) That means the density of Z at any two points z and $-z$ is the same, which means that, in computing $E(Z^3)$, positive values z^3 when $z > 0$ are exactly offset with the negative value $(-z)^3 = -z^3$. It follows that, if X is symmetric about zero, then $E(Z) = 0$. Generally, $E[(X - \mu)^3]/\sigma^3$ is viewed as a measure of **skewness** in the distribution of X. In a statistical setting, we might use data to estimate $E(Z^3)$ to determine whether an underlying population distribution appears to be symmetric. (Computer Exercise C4 in Chapter 5 provides an illustration.)

It also can be informative to compute the fourth moment of Z,

$$E(Z^4) = E[(X - \mu)^4]/\sigma^4.$$

Because $Z^4 \geq 0$, $E(Z^4) \geq 0$ (and, in any interesting case, strictly greater than zero). Without having a reference value, it is difficult to interpret values of $E(Z^4)$, but larger values mean that the tails in the distribution of X are thicker. The fourth moment $E(Z^4)$ is called a measure of **kurtosis** in the distribution of X. In Section B-5, we will obtain $E(Z^4)$ for the normal distribution.

B-4 Features of Joint and Conditional Distributions

B-4a Measures of Association: Covariance and Correlation

While the joint pdf of two random variables completely describes the relationship between them, it is useful to have summary measures of how, on average, two random variables vary with one another. As with the expected value and variance, this is similar to using a single number to summarize something about an entire distribution, which in this case is a joint distribution of two random variables.

B-4b Covariance

Let $\mu_X = E(X)$ and $\mu_Y = E(Y)$ and consider the random variable $(X - \mu_X)(Y - \mu_Y)$. Now, if X is above its mean and Y is above its mean, then $(X - \mu_X)(Y - \mu_Y) > 0$. This is also true if $X < \mu_X$ and $Y < \mu_Y$. On the other hand, if $X > \mu_X$ and $Y < \mu_Y$, or vice versa, then $(X - \mu_X)(Y - \mu_Y) < 0$. How, then, can this product tell us anything about the relationship between X and Y?

The **covariance** between two random variables X and Y, sometimes called the *population covariance* to emphasize that it concerns the relationship between two variables describing a population, is defined as the expected value of the product $(X - \mu_X)(Y - \mu_Y)$:

$$\text{Cov}(X, Y) \equiv \text{E}[(X - \mu_X)(Y - \mu_Y)], \qquad \text{[B.26]}$$

which is sometimes denoted σ_{XY}. If $\sigma_{XY} > 0$, then, on average, when X is above its mean, Y is also above its mean. If $\sigma_{XY} < 0$, then, on average, when X is above its mean, Y is below its mean.

Several expressions useful for computing $\text{Cov}(X, Y)$ are as follows:

$$\begin{aligned}\text{Cov}(X, Y) &= \text{E}[(X - \mu_X)(Y - \mu_Y)] = \text{E}[(X - \mu_X)Y] \\ &= \text{E}[X(Y - \mu_Y)] = \text{E}(XY) - \mu_X\mu_Y. \qquad \text{[B.27]}\end{aligned}$$

It follows from (B.27), that if $\text{E}(X) = 0$ or $\text{E}(Y) = 0$, then $\text{Cov}(X, Y) = \text{E}(XY)$.

Covariance measures the amount of *linear* dependence between two random variables. A positive covariance indicates that two random variables move in the same direction, while a negative covariance indicates they move in opposite directions. Interpreting the *magnitude* of a covariance can be a little tricky, as we will see shortly.

Because covariance is a measure of how two random variables are related, it is natural to ask how covariance is related to the notion of independence. This is given by the following property.

Property COV.1: If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

This property follows from equation (B.27) and the fact that $\text{E}(XY) = \text{E}(X)\text{E}(Y)$ when X and Y are independent. It is important to remember that the converse of COV.1 is *not* true: zero covariance between X and Y does not imply that X and Y are independent. In fact, there are random variables X such that, if $Y = X^2$, $\text{Cov}(X, Y) = 0$. [Any random variable with $\text{E}(X) = 0$ and $\text{E}(X^3) = 0$ has this property.] If $Y = X^2$, then X and Y are clearly not independent: once we know X, we know Y. It seems rather strange that X and X^2 could have zero covariance, and this reveals a weakness of covariance as a general measure of association between random variables. The covariance is useful in contexts when relationships are at least approximately linear.

The second major property of covariance involves covariances between linear functions.

Property COV.2: For any constants a_1, b_1, a_2, and b_2,

$$\text{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\text{Cov}(X, Y). \qquad \text{[B.28]}$$

An important implication of COV.2 is that the covariance between two random variables can be altered simply by multiplying one or both of the random variables by a constant. This is important in economics because monetary variables, inflation rates, and so on can be defined with different units of measurement without changing their meaning.

Finally, it is useful to know that the absolute value of the covariance between any two random variables is bounded by the product of their standard deviations; this is known as the *Cauchy-Schwartz inequality*.

Property COV.3: $|\text{Cov}(X, Y)| \leq \text{sd}(X)\text{sd}(Y)$.

B-4c Correlation Coefficient

Suppose we want to know the relationship between amount of education and annual earnings in the working population. We could let X denote education and Y denote earnings and then compute their covariance. But the answer we get will depend on how we choose to measure education and

earnings. Property COV.2 implies that the covariance between education and earnings depends on whether earnings are measured in dollars or thousands of dollars, or whether education is measured in months or years. It is pretty clear that how we measure these variables has no bearing on how strongly they are related. But the covariance between them does depend on the units of measurement.

The fact that the covariance depends on units of measurement is a deficiency that is overcome by the **correlation coefficient** between X and Y:

$$\text{Corr}(X, Y) \equiv \frac{\text{Cov}(X, Y)}{\text{sd}(X)\cdot\text{sd}(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}; \qquad \text{[B.29]}$$

the correlation coefficient between X and Y is sometimes denoted ρ_{XY} (and is sometimes called the *population correlation*).

Because σ_X and σ_Y are positive, $\text{Cov}(X, Y)$ and $\text{Corr}(X, Y)$ always have the same sign, and $\text{Corr}(X, Y) = 0$ if, and only if, $\text{Cov}(X, Y) = 0$. Some of the properties of covariance carry over to correlation. If X and Y are independent, then $\text{Corr}(X, Y) = 0$, but zero correlation does not imply independence. (Like the covariance, the correlation coefficient is also a measure of linear dependence.) However, the magnitude of the correlation coefficient is easier to interpret than the size of the covariance due to the following property.

Property CORR.1: $-1 \leq \text{Corr}(X, Y) \leq 1$.

If $\text{Corr}(X, Y) = 0$, or equivalently $\text{Cov}(X, Y) = 0$, then there is no linear relationship between X and Y, and X and Y are said to be **uncorrelated random variables**; otherwise, X and Y are *correlated.* $\text{Corr}(X, Y) = 1$ implies a perfect positive linear relationship, which means that we can write $Y = a + bX$ for some constant a and some constant $b > 0$. $\text{Corr}(X, Y) = -1$ implies a perfect negative linear relationship, so that $Y = a + bX$ for some $b < 0$. The extreme cases of positive or negative 1 rarely occur. Values of ρ_{XY} closer to 1 or -1 indicate stronger linear relationships.

As mentioned earlier, the correlation between X and Y is invariant to the units of measurement of either X or Y. This is stated more generally as follows.

Property CORR.2: For constants a_1, b_1, a_2, and b_2, with $a_1a_2 > 0$,

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = \text{Corr}(X, Y).$$

If $a_1a_2 < 0$, then

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = -\text{Corr}(X, Y).$$

As an example, suppose that the correlation between earnings and education in the working population is .15. This measure does not depend on whether earnings are measured in dollars, thousands of dollars, or any other unit; it also does not depend on whether education is measured in years, quarters, months, and so on.

B-4d Variance of Sums of Random Variables

Now that we have defined covariance and correlation, we can complete our list of major properties of the variance.

Property VAR.3: For constants a and b,

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y).$$

It follows immediately that, if X and Y are uncorrelated—so that $\text{Cov}(X, Y) = 0$—then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \qquad \text{[B.30]}$$

and

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y). \qquad \text{[B.31]}$$

In the latter case, note how the variance of the difference is the *sum of the variances*, not the difference in the variances.

As an example of (B.30), let X denote profits earned by a restaurant during a Friday night and let Y be profits earned on the following Saturday night. Then, $Z = X + Y$ is profits for the two nights. Suppose X and Y each have an expected value of \$300 and a standard deviation of \$15 (so that the variance is 225). Expected profits for the two nights is $\text{E}(Z) = \text{E}(X) + \text{E}(Y) = 2 \cdot (300) = 600$ dollars. If X and Y are independent, and therefore uncorrelated, then the variance of total profits is the sum of the variances: $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) = 2 \cdot (225) = 450$. It follows that the standard deviation of total profits is $\sqrt{450}$ or about \$21.21.

Expressions (B.30) and (B.31) extend to more than two random variables. To state this extension, we need a definition. The random variables $\{X_1, \ldots, X_n\}$ are **pairwise uncorrelated random variables** if each variable in the set is uncorrelated with every other variable in the set. That is, $\text{Cov}(X_i, X_j) = 0$, for all $i \neq j$.

Property VAR.4: If $\{X_1, \ldots, X_n\}$ are pairwise uncorrelated random variables and a_i: $i = 1, \ldots, n$ are constants, then

$$\text{Var}(a_1X_1 + \cdots + a_nX_n) = a_1^2\text{Var}(X_1) + \cdots + a_n^2\text{Var}(X_n).$$

In summation notation, we can write

$$\text{Var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\text{Var}(X_i). \qquad \text{[B.32]}$$

A special case of Property VAR.4 occurs when we take $a_i = 1$ for all i. Then, for pairwise uncorrelated random variables, the variance of the sum is the sum of the variances:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i). \qquad \text{[B.33]}$$

Because independent random variables are uncorrelated (see Property COV.1), the variance of a sum of independent random variables is the sum of the variances.

If the X_i are not pairwise uncorrelated, then the expression for $\text{Var}(\sum_{i=1}^{n} a_iX_i)$ is much more complicated; we must add to the right-hand side of (B.32) the terms $2a_ia_j\text{Cov}(x_i, x_j)$ for all $i > j$.

We can use (B.33) to derive the variance for a binomial random variable. Let $X \sim \text{Binomial}(n, \theta)$ and write $X = Y_1 + \cdots + Y_n$, where the Y_i are independent Bernoulli (θ) random variables. Then, by (B.33), $\text{Var}(X) = \text{Var}(Y_1) + \cdots + \text{Var}(Y_n) = n\theta(1 - \theta)$.

In the airline reservation example with $n = 120$ and $\theta = .85$, the variance of the number of passengers arriving for their reservations is $120(.85)(.15) = 15.3$, so the standard deviation is about 3.9.

B-4e Conditional Expectation

Covariance and correlation measure the linear relationship between two random variables and treat them symmetrically. More often in the social sciences, we would like to explain one variable, called Y, in terms of another variable, say, X. Further, if Y is related to X in a nonlinear fashion, we would like

to know this. Call Y the explained variable and X the explanatory variable. For example, Y might be hourly wage, and X might be years of formal education.

We have already introduced the notion of the conditional probability density function of Y given X. Thus, we might want to see how the distribution of wages changes with education level. However, we usually want to have a simple way of summarizing this distribution. A single number will no longer suffice, because the distribution of Y given $X = x$ generally depends on the value of x. Nevertheless, we can summarize the relationship between Y and X by looking at the **conditional expectation** of Y given X, sometimes called the *conditional mean.* The idea is this. Suppose we know that X has taken on a particular value, say, x. Then, we can compute the expected value of Y, given that we know this outcome of X. We denote this expected value by $\mathrm{E}(Y|X = x)$, or sometimes $\mathrm{E}(Y|x)$ for shorthand. Generally, as x changes, so does $\mathrm{E}(Y|x)$.

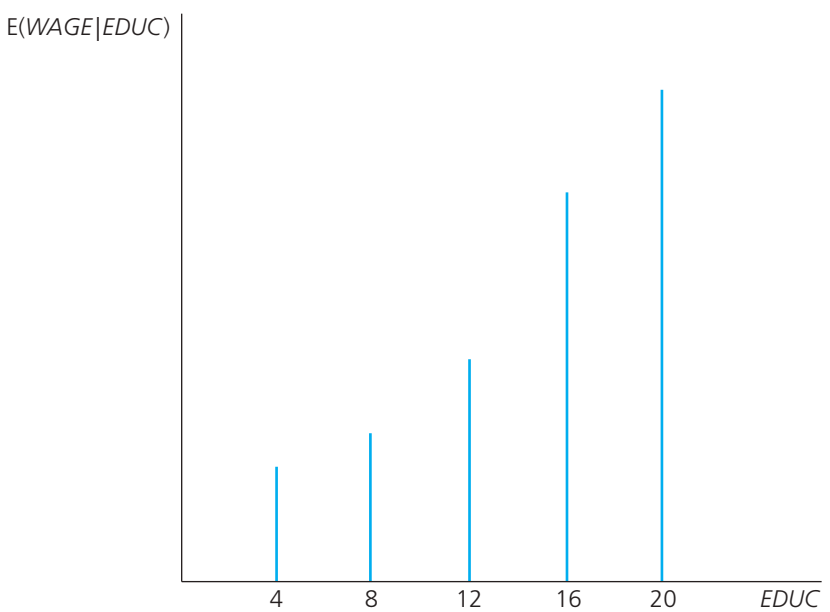
When Y is a discrete random variable taking on values $\{y_1, \ldots, y_m\}$, then

$$\mathrm{E}(Y|x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x).$$

When Y is continuous, $\mathrm{E}(Y|x)$ is defined by integrating $yf_{Y|X}(y|x)$ over all possible values of y. As with unconditional expectations, the conditional expectation is a weighted average of possible values of Y, but now the weights reflect the fact that X has taken on a specific value. Thus, $E(Y|x)$ is just some function of x, which tells us how the expected value of Y varies with x.

As an example, let (X, Y) represent the population of all working individuals, where X is years of education and Y is hourly wage. Then, $\mathrm{E}(Y|X = 12)$ is the average hourly wage for all people in the population with 12 years of education (roughly a high school education). $\mathrm{E}(Y|X = 16)$ is the average hourly wage for all people with 16 years of education. Tracing out the expected value for various levels of education provides important information on how wages and education are related. See Figure B.5 for an illustration.

FIGURE B.5 The expected value of hourly wage given various levels of education.



In principle, the expected value of hourly wage can be found at each level of education, and these expectations can be summarized in a table. Because education can vary widely—and can even be measured in fractions of a year—this is a cumbersome way to show the relationship between average wage and amount of education. In econometrics, we typically specify simple functions that capture this relationship. As an example, suppose that the expected value of *WAGE* given *EDUC* is the linear function

$$\mathrm{E}(WAGE|EDUC) = 1.05 + .45\,EDUC.$$

If this relationship holds in the population of working people, the average wage for people with eight years of education is $1.05 + .45(8) = 4.65$, or \$4.65. The average wage for people with 16 years of education is 8.25, or \$8.25. The coefficient on *EDUC* implies that each year of education increases the expected hourly wage by .45, or 45¢.

Conditional expectations can also be nonlinear functions. For example, suppose that $\mathrm{E}(Y|x) = 10/x$, where X is a random variable that is always greater than zero. This function is graphed in Figure B.6. This could represent a demand function, where Y is quantity demanded and X is price. If Y and X are related in this way, an analysis of linear association, such as correlation analysis, would be incomplete.

B-4f Properties of Conditional Expectation

Several basic properties of conditional expectations are useful for derivations in econometric analysis.

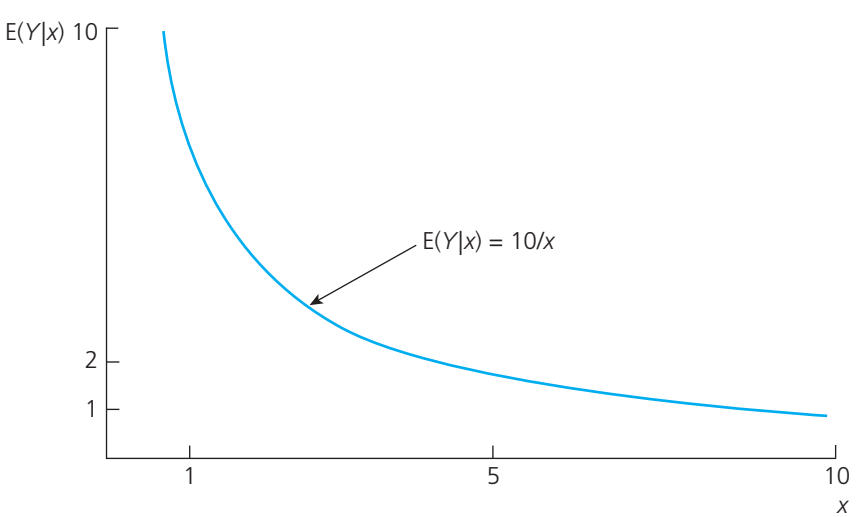
Property CE.1: $\mathrm{E}[c(X)|X] = c(X)$, for any function $c(X)$.

This first property means that functions of X behave as constants when we compute expectations conditional on X. For example, $\mathrm{E}(X^2|X) = X^2$. Intuitively, this simply means that if we know X, then we also know X^2.

Property CE.2: For functions $a(X)$ and $b(X)$,

$$\mathrm{E}[a(X)Y + b(X)|X] = a(X)\mathrm{E}(Y|X) + b(X).$$

FIGURE B.6 Graph of $\mathrm{E}(Y|x) = 10/x$.



For example, we can easily compute the conditional expectation of a function such as $XY + 2X^2$: $\mathrm{E}(XY + 2X^2|X) = X\mathrm{E}(Y|X) + 2X^2$.

The next property ties together the notions of independence and conditional expectations.

Property CE.3: If X and Y are independent, then $\mathrm{E}(Y|X) = \mathrm{E}(Y)$.

This property means that, if X and Y are independent, then the expected value of Y given X does not depend on X, in which case, $\mathrm{E}(Y|X)$ always equals the (unconditional) expected value of Y. In the wage and education example, if wages were independent of education, then the average wages of high school and college graduates would be the same. Because this is almost certainly false, we cannot assume that wage and education are independent.

A special case of Property CE.3 is the following: if U and X are independent and $\mathrm{E}(U) = 0$, then $\mathrm{E}(U|X) = 0$.

There are also properties of the conditional expectation that have to do with the fact that $\mathrm{E}(Y|X)$ is a function of X, say, $\mathrm{E}(Y|X) = \mu(X)$. Because X is a random variable, $\mu(X)$ is also a random variable. Furthermore, $\mu(X)$ has a probability distribution and therefore an expected value. Generally, the expected value of $\mu(X)$ could be very difficult to compute directly. The **law of iterated expectations** says that the expected value of $\mu(X)$ is simply equal to the expected value of Y. We write this as follows.

Property CE.4: $\mathrm{E}[\mathrm{E}(Y|X)] = \mathrm{E}(Y)$.

This property is a little hard to grasp at first. It means that, if we first obtain $\mathrm{E}(Y|X)$ as a function of X and take the expected value of this (with respect to the distribution of X, of course), then we end up with $\mathrm{E}(Y)$. This is hardly obvious, but it can be derived using the definition of expected values.

As an example of how to use Property CE.4, let $Y = WAGE$ and $X = EDUC$, where $WAGE$ is measured in hours and $EDUC$ is measured in years. Suppose the expected value of $WAGE$ given $EDUC$ is $\mathrm{E}(WAGE|EDUC) = 4 + .60\,EDUC$. Further, $\mathrm{E}(EDUC) = 11.5$. Then, the law of iterated expectations implies that $\mathrm{E}(WAGE) = \mathrm{E}(4 + .60\,EDUC) = 4 + .60\,\mathrm{E}(EDUC) = 4 + .60(11.5) = 10.90$, or \$10.90 an hour.

The next property states a more general version of the law of iterated expectations.

Property CE.4': $\mathrm{E}(Y|X) = \mathrm{E}[\mathrm{E}(Y|X, Z)|X]$.

In other words, we can find $\mathrm{E}(Y|X)$ in two steps. First, find $\mathrm{E}(Y|X, Z)$ for any other random variable Z. Then, find the expected value of $\mathrm{E}(Y|X, Z)$, conditional on X.

Property CE.5: If $\mathrm{E}(Y|X) = \mathrm{E}(Y)$, then $\mathrm{Cov}(X, Y) = 0$ [and so $\mathrm{Corr}(X, Y) = 0$]. In fact, *every* function of X is uncorrelated with Y.

This property means that, if knowledge of X does not change the expected value of Y, then X and Y *must* be uncorrelated, which implies that if X and Y are correlated, then $\mathrm{E}(Y|X)$ must depend on X. The converse of Property CE.5 is not true: if X and Y are uncorrelated, $\mathrm{E}(Y|X)$ *could* still depend on X. For example, suppose $Y = X^2$. Then, $\mathrm{E}(Y|X) = X^2$, which is clearly a function of X. However, as we mentioned in our discussion of covariance and correlation, it is possible that X and X^2 are uncorrelated. The conditional expectation captures the nonlinear relationship between X and Y that correlation analysis would miss entirely.

Properties CE.4 and CE.5 have two important implications: if U and X are random variables such that $\mathrm{E}(U|X) = 0$, then $\mathrm{E}(U) = 0$, and U and X are uncorrelated.

Property CE.6: If $\mathrm{E}(Y^2) < \infty$ and $\mathrm{E}[g(X)^2] < \infty$ for some function g, then $\mathrm{E}\{[Y - \mu(X)]^2|X\} \leq \mathrm{E}\{[Y - g(X)]^2|X\}$ and $\mathrm{E}\{[Y - \mu(X)]^2\} \leq \mathrm{E}\{[Y - g(X)]^2\}$.

Property CE.6 is very useful in predicting or forecasting contexts. The first inequality says that, if we measure prediction inaccuracy as the *expected* squared prediction error, conditional on X, then the conditional mean is better than any other function of X for predicting Y. The conditional mean also minimizes the unconditional expected squared prediction error.

B-4g Conditional Variance

Given random variables X and Y, the variance of Y, conditional on $X = x$, is simply the variance associated with the conditional distribution of Y, given $X = x$: $\mathrm{E}\{[Y - \mathrm{E}(Y|x)]^2|x\}$. The formula

$$\mathrm{Var}(Y|X = x) = \mathrm{E}(Y^2|x) - [\mathrm{E}(Y|x)]^2$$

is often useful for calculations. Only occasionally will we have to compute a conditional variance. But we will have to make assumptions about and manipulate conditional variances for certain topics in regression analysis.

As an example, let $Y = SAVING$ and $X = INCOME$ (both of these measured annually for the population of all families). Suppose that $\mathrm{Var}(SAVING|INCOME) = 400 + .25\,INCOME$. This says that, as income increases, the variance in saving levels also increases. It is important to see that the relationship between the variance of *SAVING* and *INCOME* is totally separate from that between the *expected value* of *SAVING* and *INCOME*.

We state one useful property about the conditional variance.

Property CV.1: If X and Y are independent, then $\mathrm{Var}(Y|X) = \mathrm{Var}(Y)$.

This property is pretty clear, as the distribution of Y given X does not depend on X, and $\mathrm{Var}(Y|X)$ is just one feature of this distribution.

B-5 The Normal and Related Distributions

B-5a The Normal Distribution

The normal distribution and those derived from it are the most widely used distributions in statistics and econometrics. Assuming that random variables defined over populations are normally distributed simplifies probability calculations. In addition, we will rely heavily on the normal and related distributions to conduct inference in statistics and econometrics—even when the underlying population is not necessarily normal. We must postpone the details, but be assured that these distributions will arise many times throughout this text.

A normal random variable is a continuous random variable that can take on any value. Its probability density function has the familiar bell shape graphed in Figure B.7.

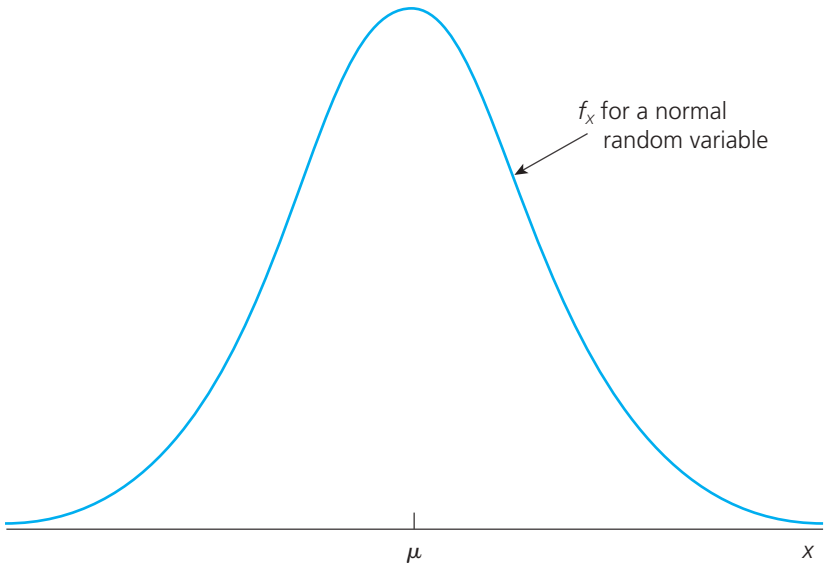
Mathematically, the pdf of X can be written as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-(x - \mu)^2/2\sigma^2], \quad -\infty < x < \infty, \qquad [B.34]$$

where $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$. We say that X has a **normal distribution** with expected value μ and variance σ^2, written as $X \sim \mathrm{Normal}(\mu, \sigma^2)$. Because the normal distribution is symmetric about μ, μ is also the median of X. The normal distribution is sometimes called the *Gaussian distribution* after the famous mathematician C. F. Gauss.

Certain random variables appear to roughly follow a normal distribution. Human heights and weights, test scores, and county unemployment rates have pdfs roughly the shape in Figure B.7. Other distributions, such as income distributions, do not appear to follow the normal probability function. In most countries, income is not symmetrically distributed about any value; the distribution is skewed toward the upper tail. In some cases, a variable can be transformed to achieve normality. A popular transformation is

FIGURE B.7 The general shape of the normal probability density function.



the natural log, which makes sense for positive random variables. If X is a positive random variable, such as income, and $Y = \log(X)$ has a normal distribution, then we say that X has a *lognormal distribution*. It turns out that the lognormal distribution fits income distribution pretty well in many countries. Other variables, such as prices of goods, appear to be well described as lognormally distributed.

B-5b The Standard Normal Distribution

One special case of the normal distribution occurs when the mean is zero and the variance (and, therefore, the standard deviation) is unity. If a random variable Z has a Normal(0,1) distribution, then we say it has a **standard normal distribution**. The pdf of a standard normal random variable is denoted $\phi(z)$; from (B.34), with $\mu = 0$ and $\sigma^2 = 1$, it is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2), \quad -\infty < z < \infty. \qquad \text{[B.35]}$$

The standard normal cumulative distribution function is denoted $\Phi(z)$ and is obtained as the area under ϕ, to the left of z; see Figure B.8. Recall that $\Phi(z) = \mathrm{P}(Z \leq z)$; because Z is continuous, $\Phi(z) = \mathrm{P}(Z < z)$ as well.

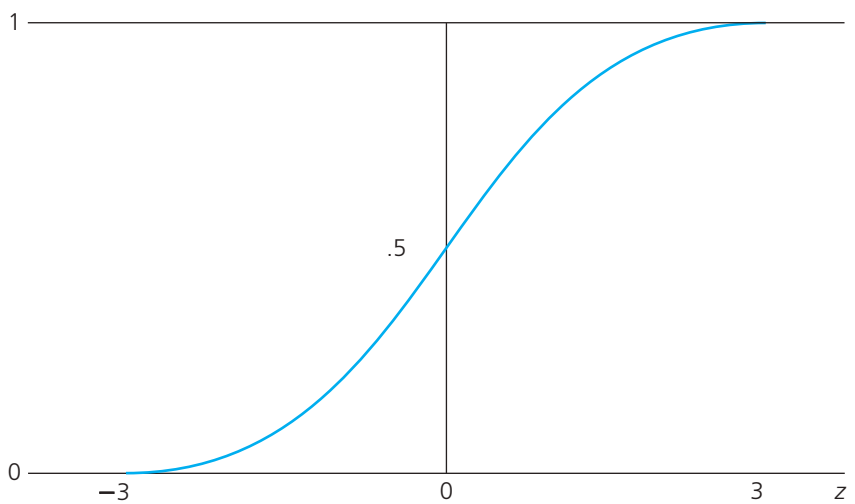
No simple formula can be used to obtain the values of $\Phi(z)$ [because $\Phi(z)$ is the integral of the function in (B.35), and this integral has no closed form]. Nevertheless, the values for $\Phi(z)$ are easily tabulated; they are given for z between −3.1 and 3.1 in Table G.1 in Statistical Tables. For $z \leq -3.1$, $\Phi(z)$ is less than .001, and for $z \geq 3.1$, $\Phi(z)$ is greater than .999. Most statistics and econometrics software packages include simple commands for computing values of the standard normal cdf, so we can often avoid printed tables entirely and obtain the probabilities for any value of z.

Using basic facts from probability—and, in particular, properties (B.7) and (B.8) concerning cdfs—we can use the standard normal cdf for computing the probability of any event involving a standard normal random variable. The most important formulas are

$$\mathrm{P}(Z > z) = 1 - \Phi(z), \qquad \text{[B.36]}$$

$$\mathrm{P}(Z < -z) = \mathrm{P}(Z > z), \qquad \text{[B.37]}$$

FIGURE B.8 The standard normal cumulative distribution function.



and

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a). \qquad [B.38]$$

Because Z is a continuous random variable, all three formulas hold whether or not the inequalities are strict. Some examples include $P(Z > .44) = 1 - .67 = .33$, $P(Z < -.92) = P(Z > .92) = 1 - .821 = .179$, and $P(-1 < Z \leq .5) = .692 - .159 = .533$.

Another useful expression is that, for any $c > 0$,

$$\begin{aligned} P(|Z| > c) &= P(Z > c) + P(Z < -c) \\ &= 2 \cdot P(Z > c) = 2[1 - \Phi(c)]. \end{aligned} \qquad [B.39]$$

Thus, the probability that the absolute value of Z is bigger than some positive constant c is simply twice the probability $P(Z > c)$; this reflects the symmetry of the standard normal distribution.

In most applications, we start with a normally distributed random variable, $X \sim \text{Normal}(\mu, \sigma^2)$, where μ is different from zero and $\sigma^2 \neq 1$. Any normal random variable can be turned into a standard normal using the following property.

Property Normal.1: If $X \sim \text{Normal}(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim \text{Normal}(0, 1)$.

Property Normal.1 shows how to turn any normal random variable into a standard normal. Thus, suppose $X \sim \text{Normal}(3, 4)$, and we would like to compute $P(X \leq 1)$. The steps always involve the normalization of X to a standard normal:

$$\begin{aligned} P(X \leq 1) &= P(X - 3 \leq 1 - 3) = P\left(\frac{X - 3}{2} \leq -1\right) \\ &= P(Z \leq -1) = \Phi(-1) = .159. \end{aligned}$$

EXAMPLE B.6 **Probabilities for a Normal Random Variable**

First, let us compute $P(2 < X \leq 6)$ when $X \sim \text{Normal}(4,9)$ (whether we use $<$ or $\leq$ is irrelevant because X is a continuous random variable). Now,

$$P(2 < X \leq 6) = P\left(\frac{2-4}{3} < \frac{X-4}{3} \leq \frac{6-4}{3}\right) = P(-2/3 < Z \leq 2/3)$$
$$= \Phi(.67) - \Phi(-.67) = .749 - .251 = .498.$$

Now, let us compute $P(|X| > 2)$:

$$P(|X| > 2) = P(X > 2) + P(X < -2)$$
$$= P[(X-4)/3 > (2-4)/3] + P[(X-4)/3 < (-2-4)/3]$$
$$= 1 - \Phi(-2/3) + \Phi(-2)$$
$$= 1 - .251 + .023 = .772.$$

B-5c Additional Properties of the Normal Distribution

We end this subsection by collecting several other facts about normal distributions that we will later use.

Property Normal.2: If $X \sim \text{Normal}(\mu, \sigma^2)$, then $aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$.

Thus, if $X \sim \text{Normal}(1,9)$, then $Y = 2X + 3$ is distributed as normal with mean $2E(X) + 3 = 5$ and variance $2^2 \cdot 9 = 36$; $\text{sd}(Y) = 2\text{sd}(X) = 2 \cdot 3 = 6$.

Earlier, we discussed how, in general, zero correlation and independence are not the same. In the case of normally distributed random variables, it turns out that zero correlation suffices for independence.

Property Normal.3: If X and Y are jointly normally distributed, then they are independent if, and only if, $\text{Cov}(X, Y) = 0$.

Property Normal.4: Any linear combination of independent, identically distributed normal random variables has a normal distribution.

For example, let X_i, for $i = 1, 2$, and 3, be independent random variables distributed as $\text{Normal}(\mu, \sigma^2)$. Define $W = X_1 + 2X_2 - 3X_3$. Then, W is normally distributed; we must simply find its mean and variance. Now,

$$E(W) = E(X_1) + 2E(X_2) - 3E(X_3) = \mu + 2\mu - 3\mu = 0.$$

Also,

$$\text{Var}(W) = \text{Var}(X_1) + 4\text{Var}(X_2) + 9\text{Var}(X_3) = 14\sigma^2.$$

Property Normal.4 also implies that the average of independent, normally distributed random variables has a normal distribution. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables and each is distributed as $\text{Normal}(\mu, \sigma^2)$, then

$$\overline{Y} \sim \text{Normal}(\mu, \sigma^2/n). \qquad \text{[B.40]}$$

This result is critical for statistical inference about the mean in a normal population.

Other features of the normal distribution are worth knowing, although they do not play a central role in the text. Because a normal random variable is symmetric about its mean, it has zero skewness, that is, $E[(X - \mu)^3] = 0$. Further, it can be shown that

$$E[(X - \mu)^4]/\sigma^4 = 3,$$

or $E(Z^4) = 3$, where Z has a standard normal distribution. Because the normal distribution is so prevalent in probability and statistics, the measure of kurtosis for any given random variable X (whose fourth moment exists) is often defined to be $E[(X - \mu)^4]/\sigma^4 - 3$, that is, relative to the value for the standard normal distribution. If $E[(X - \mu)^4]/\sigma^4 > 3$, then the distribution of X has fatter tails than the normal distribution (a somewhat common occurrence, such as with the t distribution to be introduced shortly); if $E[(X - \mu)^4]/\sigma^4 < 3$, then the distribution has thinner tails than the normal (a rarer situation).

B-5d The Chi-Square Distribution

The chi-square distribution is obtained directly from independent, standard normal random variables. Let Z_i, $i = 1, 2, \ldots, n$, be independent random variables, each distributed as standard normal. Define a new random variable as the sum of the squares of the Z_i:

$$X = \sum_{i=1}^{n} Z_i^2. \qquad [B.41]$$

Then, X has what is known as a **chi-square distribution** with n **degrees of freedom** (or *df* for short). We write this as $X \sim \chi_n^2$. The *df* in a chi-square distribution corresponds to the number of terms in the sum in (B.41). The concept of degrees of freedom will play an important role in our statistical and econometric analyses.

The pdf for chi-square distributions with varying degrees of freedom is given in Figure B.9; we will not need the formula for this pdf, and so we do not reproduce it here. From equation (B.41), it is clear that a chi-square random variable is always nonnegative, and that, unlike the normal distribution, the chi-square distribution is not symmetric about any point. It can be shown that if $X \sim \chi_n^2$, then the expected value of X is n [the number of terms in (B.41)], and the variance of X is $2n$.

B-5e The *t* Distribution

The t distribution is the workhorse in classical statistics and multiple regression analysis. We obtain a t distribution from a standard normal and a chi-square random variable.

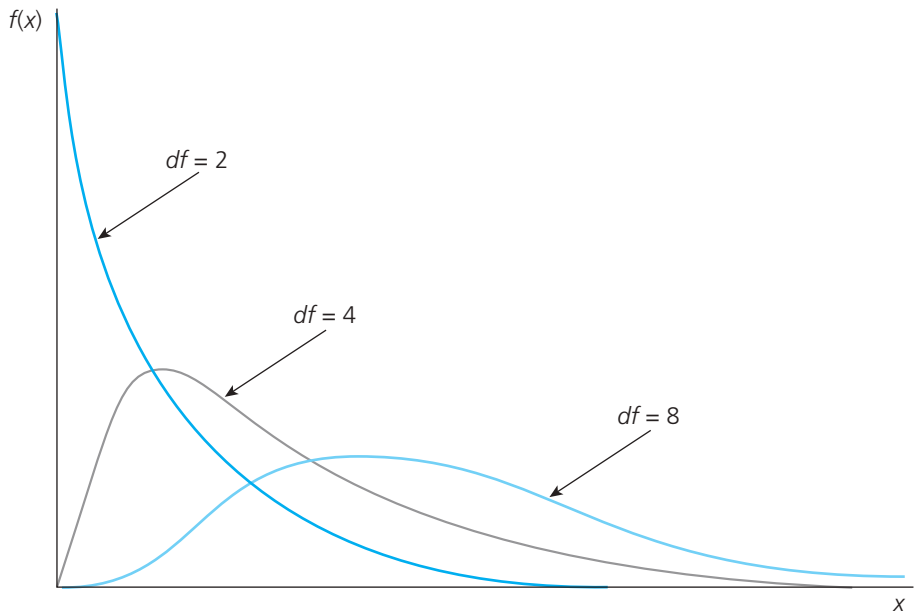
Let Z have a standard normal distribution and let X have a chi-square distribution with n degrees of freedom. Further, assume that Z and X are independent. Then, the random variable

$$T = \frac{Z}{\sqrt{X/n}} \qquad [B.42]$$

has a ***t* distribution** with n degrees of freedom. We will denote this by $T \sim t_n$. The t distribution gets its degrees of freedom from the chi-square random variable in the denominator of (B.42).

The pdf of the t distribution has a shape similar to that of the standard normal distribution, except that it is more spread out and therefore has more area in the tails. The expected value of a t distributed random variable is zero (strictly speaking, the expected value exists only for $n > 1$),

FIGURE B.9 The chi-square distribution with various degrees of freedom.



and the variance is $n/(n-2)$ for $n > 2$. (The variance does not exist for $n \leq 2$ because the distribution is so spread out.) The pdf of the t distribution is plotted in Figure B.10 for various degrees of freedom. As the degrees of freedom gets large, the t distribution approaches the standard normal distribution.

B-5f The *F* Distribution

Another important distribution for statistics and econometrics is the F distribution. In particular, the F distribution will be used for testing hypotheses in the context of multiple regression analysis.

To define an F random variable, let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ and assume that X_1 and X_2 are independent. Then, the random variable

$$F = \frac{(X_1/k_1)}{(X_2/k_2)} \qquad [B.43]$$

has an ***F* distribution** with (k_1, k_2) degrees of freedom. We denote this as $F \sim F_{k_1, k_2}$. The pdf of the F distribution with different degrees of freedom is given in Figure B.11.

The order of the degrees of freedom in F_{k_1, k_2} is critical. The integer k_1 is called the *numerator degrees of freedom* because it is associated with the chi-square variable in the numerator. Likewise, the integer k_2 is called the *denominator degrees of freedom* because it is associated with the chi-square variable in the denominator. This can be a little tricky because (B.43) can also be written as $(X_1k_2)/(X_2k_1)$, so that k_1 appears in the denominator. Just remember that the numerator *df* is the integer associated with the chi-square variable in the numerator of (B.43), and similarly for the denominator *df*.

FIGURE B.10 The *t* distribution with various degrees of freedom.

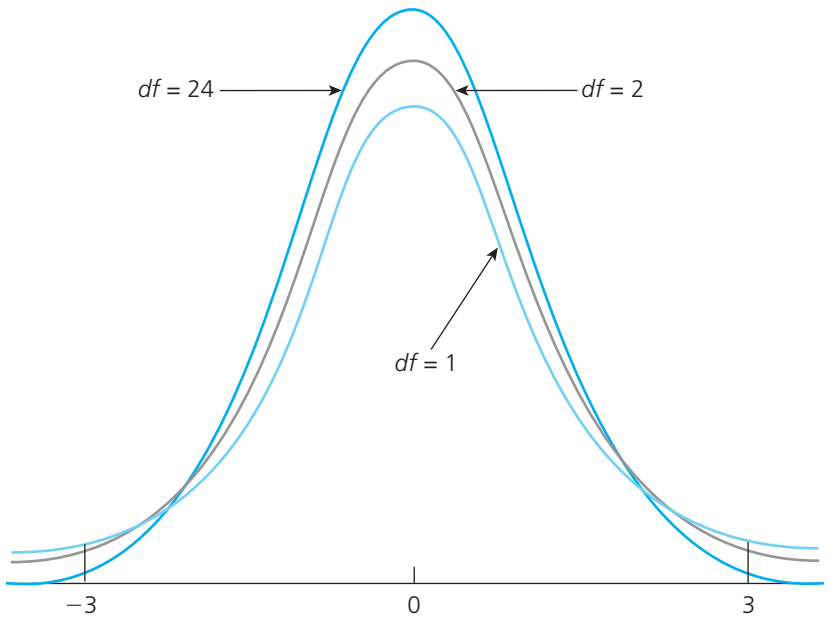
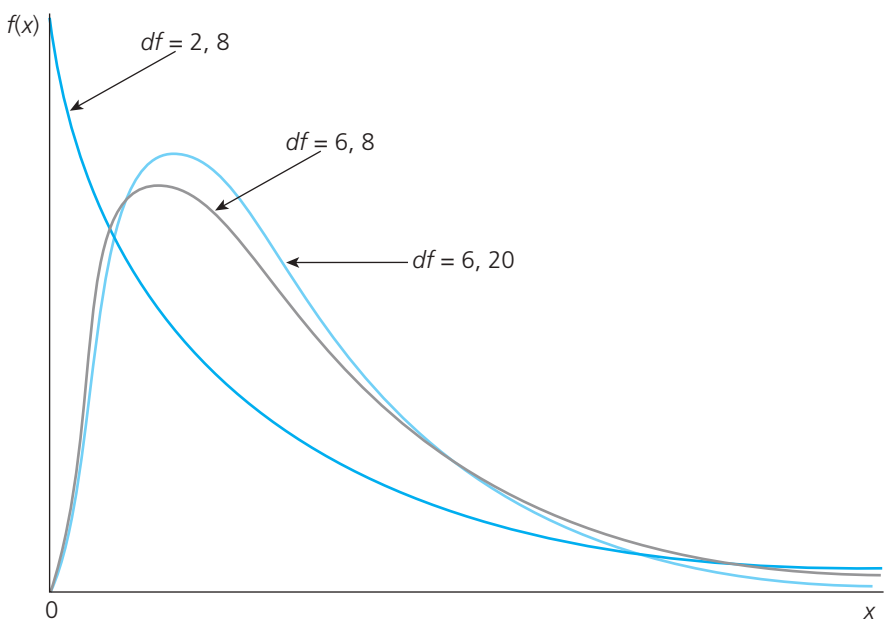


FIGURE B.11 The F_{k_1, k_2} distribution for various degrees of freedom, k_1 and k_2.



Summary

In this Math Refresher, we have reviewed the probability concepts that are needed in econometrics. Most of the concepts should be familiar from your introductory course in probability and statistics. Some of the more advanced topics, such as features of conditional expectations, do not need to be mastered now—there is time for that when these concepts arise in the context of regression analysis in Part 1.

In an introductory statistics course, the focus is on calculating means, variances, covariances, and so on for particular distributions. In Part 1, we will not need such calculations: we mostly rely on the *properties* of expectations, variances, and so on that have been stated in this Math Refresher.

Key Terms

Bernoulli (or Binary) Random Variable
Binomial Distribution
Chi-Square Distribution
Conditional Distribution
Conditional Expectation
Continuous Random Variable
Correlation Coefficient
Covariance
Cumulative Distribution Function (cdf)
Degrees of Freedom
Discrete Random Variable
Expected Value
Experiment
F Distribution
Independent Random Variables
Joint Distribution
Kurtosis
Law of Iterated Expectations
Median
Normal Distribution
Pairwise Uncorrelated Random Variables
Probability Density Function (pdf)
Random Variable
Skewness
Standard Deviation
Standard Normal Distribution
Standardized Random Variable
Symmetric Distribution
t Distribution
Uncorrelated Random Variables
Variance

Problems

1 Suppose that a high school student is preparing to take the SAT exam. Explain why his or her eventual SAT score is properly viewed as a random variable.

2 Let X be a random variable distributed as Normal(5,4). Find the probabilities of the following events:
(i) $P(X \leq 6)$.
(ii) $P(X > 4)$.
(iii) $P(|X - 5| > 1)$.

3 Much is made of the fact that certain mutual funds outperform the market year after year (that is, the return from holding shares in the mutual fund is higher than the return from holding a portfolio such as the S&P 500). For concreteness, consider a 10-year period and let the population be the 4,170 mutual funds reported in *The Wall Street Journal* on January 1, 1995. By saying that performance relative to the market is random, we mean that each fund has a 50–50 chance of outperforming the market in any year and that performance is independent from year to year.
(i) If performance relative to the market is truly random, what is the probability that any particular fund outperforms the market in all 10 years?
(ii) Of the 4,170 mutual funds, what is the expected number of funds that will outperform the market in all 10 years?
(iii) Find the probability that *at least* one fund out of 4,170 funds outperforms the market in all 10 years. What do you make of your answer?
(iv) If you have a statistical package that computes binomial probabilities, find the probability that at least five funds outperform the market in all 10 years.

4 For a randomly selected county in the United States, let X represent the proportion of adults over age 65 who are employed, or the elderly employment rate. Then, X is restricted to a value between zero and one. Suppose that the cumulative distribution function for X is given by $F(x) = 3x^2 - 2x^3$ for $0 \leq x \leq 1$. Find the probability that the elderly employment rate is at least .6 (60%).

5 Just prior to jury selection for O. J. Simpson's murder trial in 1995, a poll found that about 20% of the adult population believed Simpson was innocent (after much of the physical evidence in the case had been revealed to the public). Ignore the fact that this 20% is an estimate based on a subsample from the population; for illustration, take it as the true percentage of people who thought Simpson was innocent prior to jury selection. Assume that the 12 jurors were selected randomly and independently from the population (although this turned out not to be true).

(i) Find the probability that the jury had at least one member who believed in Simpson's innocence prior to jury selection. [*Hint*: Define the Binomial(12,.20) random variable X to be the number of jurors believing in Simpson's innocence.]

(ii) Find the probability that the jury had at least two members who believed in Simpson's innocence. [*Hint*: $P(X \geq 2) = 1 - P(X \leq 1)$ and $P(X \leq 1) = P(X = 0) + P(X = 1)$.]

6 (Requires calculus) Let X denote the prison sentence, in years, for people convicted of auto theft in a particular state in the United States. Suppose that the pdf of X is given by

$$f(x) = (1/9)x^2, 0 < x < 3.$$

Use integration to find the expected prison sentence.

7 If a basketball player is a 74% free throw shooter, then, on average, how many free throws will he or she make in a game with eight free throw attempts?

8 Suppose that a college student is taking three courses: a two-credit course, a three-credit course, and a four-credit course. The expected grade in the two-credit course is 3.5, while the expected grade in the three- and four-credit courses is 3.0. What is the expected overall grade point average for the semester? (Remember that each course grade is weighted by its share of the total number of units.)

9 Let X denote the annual salary of university professors in the United States, measured in thousands of dollars. Suppose that the average salary is 52.3, with a standard deviation of 14.6. Find the mean and standard deviation when salary is measured in dollars.

10 Suppose that at a large university, college grade point average, *GPA*, and SAT score, *SAT*, are related by the conditional expectation $E(GPA|SAT) = .70 + .002\,SAT$.

(i) Find the expected *GPA* when $SAT = 800$. Find $E(GPA|SAT = 1{,}400)$. Comment on the difference.

(ii) If the average *SAT* in the university is 1,100, what is the average *GPA*? (*Hint*: Use Property CE.4.)

(iii) If a student's SAT score is 1,100, does this mean he or she will have the GPA found in part (ii)? Explain.

11 (i) Let X be a random variable taking on the values -1 and 1, each with probability 1/2. Find $E(X)$ and $E(X^2)$.

(ii) Now let X be a random variable taking on the values 1 and 2, each with probability 1/2. Find $E(X)$ and $E(1/X)$.

(iii) Conclude from parts (i) and (ii) that, in general,

$$E[g(X)] \neq g[E(X)]$$

for a nonlinear function $g(\cdot)$.

(iv) Given the definition of the F random variable in equation (B.43), show that

$$E(F) = E\left[\frac{1}{(X_2/k_2)}\right].$$

Can you conclude that $E(F) = 1$?

12 The *geometric distribution* can be used to model the number of trials before a certain event occurs. For example, we might flip a coin repeatedly until the first head appears. If the coin is fair, the probability of getting a head on each flip is 0. 5. Furthermore, we may realistically assume that the trials are independent. The flip on which the first head occurs can be represented by a random variable, X.

For the general geometric distribution, we maintain the assumption of independent trails—which, admittedly, is sometimes too strong—but allow the probability of the event occurring on any trial to be θ for any $0 < \theta < 1$. We assume that this probability is the same from trial to trial. In the coin-flipping example, allowing the coin to be biased toward, say, heads, would mean $\theta > 0.5$. Another example would be an unemployed worker repeatedly interviewing for jobs until the first job offer. Then θ is the probability of receiving an offer during any particular interview. To follow the geometric distribution, we would assume θ is the same for all interviews and that the outcomes are independent across interviews. Both assumptions may be too strong.

One way to characterize the geometric distribution is to define a sequence of Bernoulli (binary) variables, say $W_1, W_2, W_3, \ldots$. If $W_k = 1$ then the event occurs on trial k; if $W_k = 0$, it does not occur. Assume that the W_k are independent across k with the Bernoulli(θ) distribution, so that $P(W_k = 1) = \theta$.

(i) Let X denote the trial upon which the first event occurs. The possible values of X are $\{1, 2, 3, \ldots\}$. Show that for any positive integer k,

$$P(X = k) = (1 - \theta)^{k-1}\theta.$$

[Hint: If $X = k$, you must observe $k - 1$ "failures" (zeros) followed by a "success" (one).]

(ii) Use the formula for a geometric sum to show that

$$P(X \leq k) = 1 - (1 - \theta)^{k}, k = 1, 2, \ldots.$$

(iii) Suppose you have observed 29 failures in a row. If $\theta = 0.04$, what is the probability of observing a success on the 30^{th} trial?

(iv) In the setup of part (iii), before conducting any of the trials, what is the probability that the first success occurs before the 30^{th} trial?

(v) Reconcile your answers from parts (iii) and (iv).

13 In March of 1985, the NCAA men's basketball tournament increased its field of teams to 64. Since that time, each year of the tournament involves four games pitting a #1 seed against a #16 seed. The #1 seeds are purportedly awarded to the four most deserving teams. The #16 teams are generally viewed as the weakest four teams in the field. In answering this question, we will make some simplifying assumptions to make the calculations easier.

(i) Assume that the probability of a #16 seed beating a #1 seed is ρ, where $0 < \rho < 1$. (In practice, ρ varies by matchup, but we will assume it is the same across all matchups and years.) Assume that the outcomes of #1 vs #16 games are independent of one another. Show that the probability that at least one #16 seed wins in a particular year is $1 - (1 - \rho)^4$. Evaluate this probability when $\rho = 0.02$. [Hint: You might define four binary variables, say Z_1, Z_2, Z_3, and Z_4,where $Z_i = 1$ if game i is won by the #16 seed. Then first compute $P(Z_1 = 0, Z_2 = 0, Z_3 = 0, Z_4 = 0)$.]

(ii) Let X be the number of years before a #16 beats a #1 seed in the tournament. Assuming independence across years—a very reasonable assumption—explain why X has a geometric distribution and that the probability of "success" on a given trial is $\theta = 1 - (1 - \rho)^4$.

(iii) In the 2017 NCCA Men's Tournament, #16 seed University of Maryland, Baltimore County defeated #1 seed University of Virginia. It took 33 years for such an upset to occur. Suppose $\rho = 0.02$. Find $P(X \leq 33)$. Interpret this probability using the perspective of a basketball observer in February 1985.

(iv) Using $\rho = 0.02$, in February 2018 what was the probability that a #16 seed would defeat a #1 seed in the March 2018 tournament? (It had not happened in the previous 32 years.) Why does this differ so much from your answer in part (iii)?

(v) Derive the general formula

$$P(X \leq k) = 1 - (1 - \rho)^{4k}.$$

Math Refresher C

Fundamentals of Mathematical Statistics

C-1 Populations, Parameters, and Random Sampling

Statistical inference involves learning something about a population given the availability of a sample from that population. By **population**, we mean any well-defined group of subjects, which could be individuals, firms, cities, or many other possibilities. By "learning," we can mean several things, which are broadly divided into the categories of *estimation* and *hypothesis testing*.

A couple of examples may help you understand these terms. In the population of all working adults in the United States, labor economists are interested in learning about the return to education, as measured by the average percentage increase in earnings given another year of education. It would be impractical and costly to obtain information on earnings and education for the entire working population in the United States, but we can obtain data on a subset of the population. Using the data collected, a labor economist may report that his or her best estimate of the return to another year of education is 7.5%. This is an example of a *point estimate*. Or, she or he may report a range, such as "the return to education is between 5.6% and 9.4%." This is an example of an *interval estimate*.

An urban economist might want to know whether neighborhood crime watch programs are associated with lower crime rates. After comparing crime rates of neighborhoods with and without such programs in a sample from the population, he or she can draw one of two conclusions: neighborhood watch programs do affect crime, or they do not. This example falls under the rubric of hypothesis testing.

The first step in statistical inference is to identify the population of interest. This may seem obvious, but it is important to be very specific. Once we have identified the population, we can specify a model for the population relationship of interest. Such models involve probability distributions or features of probability distributions, and these depend on unknown parameters. Parameters are simply constants that determine the directions and strengths of relationships among variables. In the labor economics example just presented, the parameter of interest is the return to education in the population.

C-1a Sampling

For reviewing statistical inference, we focus on the simplest possible setting. Let Y be a random variable representing a population with a probability density function $f(y; \theta)$, which depends on the single parameter θ. The probability density function (pdf) of Y is assumed to be known except for the value of θ; different values of θ imply different population distributions, and therefore we are interested in the value of θ. If we can obtain certain kinds of samples from the population, then we can learn something about θ. The easiest sampling scheme to deal with is random sampling.

Random Sampling. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables with a common probability density function $f(y; \theta)$, then $\{Y_1, \ldots, Y_n\}$ is said to be a **random sample** from $f(y; \theta)$ [or a random sample from the population represented by $f(y; \theta)$].

When $\{Y_1, \ldots, Y_n\}$ is a random sample from the density $f(y; \theta)$, we also say that the Y_i are *independent, identically distributed* (or *i.i.d.*) random variables from $f(y; \theta)$. In some cases, we will not need to entirely specify what the common distribution is.

The random nature of $Y_1, Y_2, \ldots, Y_n$ in the definition of random sampling reflects the fact that many different outcomes are possible before the sampling is actually carried out. For example, if family income is obtained for a sample of $n = 100$ families in the United States, the incomes we observe will usually differ for each different sample of 100 families. Once a sample is obtained, we have a set of numbers, say, $\{y_1, y_2, \ldots, y_n\}$, which constitute the data that we work with. Whether or not it is appropriate to assume the sample came from a random sampling scheme requires knowledge about the actual sampling process.

Random samples from a Bernoulli distribution are often used to illustrate statistical concepts, and they also arise in empirical applications. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables and each is distributed as Bernoulli(θ), so that $P(Y_i = 1) = \theta$ and $P(Y_i = 0) = 1 - \theta$, then $\{Y_1, Y_2, \ldots, Y_n\}$ constitutes a random sample from the Bernoulli(θ) distribution. As an illustration, consider the airline reservation example carried along in Math Refresher B. Each Y_i denotes whether customer i shows up for his or her reservation; $Y_i = 1$ if passenger i shows up, and $Y_i = 0$ otherwise. Here, θ is the probability that a randomly drawn person from the population of all people who make airline reservations shows up for his or her reservation.

For many other applications, random samples can be assumed to be drawn from a normal distribution. If $\{Y_1, \ldots, Y_n\}$ is a random sample from the Normal(μ, σ^2) population, then the population is characterized by two parameters, the mean μ and the variance σ^2. Primary interest usually lies in μ, but σ^2 is of interest in its own right because making inferences about μ often requires learning about σ^2.

C-2 Finite Sample Properties of Estimators

In this section, we study what are called finite sample properties of estimators. The term "finite sample" comes from the fact that the properties hold for a sample of any size, no matter how large or small. Sometimes, these are called small sample properties. In Section C-3, we cover "asymptotic properties," which have to do with the behavior of estimators as the sample size grows without bound.

C-2a Estimators and Estimates

To study properties of estimators, we must define what we mean by an estimator. Given a random sample $\{Y_1, Y_2, \ldots, Y_n\}$ drawn from a population distribution that depends on an unknown parameter θ, an **estimator** of θ is a rule that assigns each possible outcome of the sample a value of θ. The rule is specified before any sampling is carried out; in particular, the rule is the same regardless of the data actually obtained.

As an example of an estimator, let $\{Y_1, \ldots, Y_n\}$ be a random sample from a population with mean μ. A natural estimator of μ is the average of the random sample:

$$\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i. \qquad \text{[C.1]}$$

$\bar{Y}$ is called the **sample average** but, unlike in Math Refresher A where we defined the sample average of a set of numbers as a descriptive statistic, $\bar{Y}$ is now viewed as an estimator. Given any outcome of the random variables $Y_1, \ldots, Y_n$, we use the same rule to estimate μ: we simply average them. For actual data outcomes $\{y_1, \ldots, y_n\}$, the **estimate** is just the average in the sample: $\bar{y} = (y_1 + y_2 + \cdots + y_n)/n$.

EXAMPLE C.1 **City Unemployment Rates**

Suppose we obtain the following sample of unemployment rates for 10 cities in the United States:

City	Unemployment Rate
1	5.1
2	6.4
3	9.2
4	4.1
5	7.5
6	8.3
7	2.6
8	3.5
9	5.8
10	7.5

Our estimate of the average city unemployment rate in the United States is $\bar{y} = 6.0$. Each sample generally results in a different estimate. But the *rule* for obtaining the estimate is the same, regardless of which cities appear in the sample, or how many.

More generally, an estimator W of a parameter θ can be expressed as an abstract mathematical formula:

$$W = h(Y_1, Y_2, \ldots, Y_n), \qquad \text{[C.2]}$$

for some known function h of the random variables $Y_1, Y_2, \ldots, Y_n$. As with the special case of the sample average, W is a random variable because it depends on the random sample: as we obtain different random samples from the population, the value of W can change. When a particular set of numbers, say, $\{y_1, y_2, \ldots, y_n\}$, is plugged into the function h, we obtain an *estimate* of θ, denoted $w = h(y_1, \ldots, y_n)$. Sometimes, W is called a point estimator and w a point estimate to distinguish these from *interval* estimators and estimates, which we will come to in Section C-5.

For evaluating estimation procedures, we study various properties of the probability distribution of the random variable W. The distribution of an estimator is often called its **sampling distribution**, because this distribution describes the likelihood of various outcomes of W across different random samples. Because there are unlimited rules for combining data to estimate parameters, we need some sensible criteria for choosing among estimators, or at least for eliminating some estimators from consideration. Therefore, we must leave the realm of descriptive statistics, where we compute things such as the sample average to simply summarize a body of data. In mathematical statistics, we study the sampling distributions of estimators.

C-2b Unbiasedness

In principle, the entire sampling distribution of W can be obtained given the probability distribution of Y_i and the function h. It is usually easier to focus on a few features of the distribution of W in evaluating it as an estimator of θ. The first important property of an estimator involves its expected value.

Unbiased Estimator. An estimator, W of θ, is an **unbiased estimator** if

$$\mathrm{E}(W) = \theta, \qquad \text{[C.3]}$$

for all possible values of θ.

If an estimator is unbiased, then its probability distribution has an expected value equal to the parameter it is supposed to be estimating. Unbiasedness does *not* mean that the estimate we get with any particular sample is equal to θ, or even very close to θ. Rather, if we could *indefinitely* draw random samples on Y from the population, compute an estimate each time, and then average these estimates over all random samples, we would obtain θ. This thought experiment is abstract because, in most applications, we just have one random sample to work with.

For an estimator that is not unbiased, we define its **bias** as follows.

Bias of an Estimator. If W is a **biased estimator** of θ, its bias is defined

$$\text{Bias}(W) \equiv \text{E}(W) - \theta. \qquad \text{[C.4]}$$

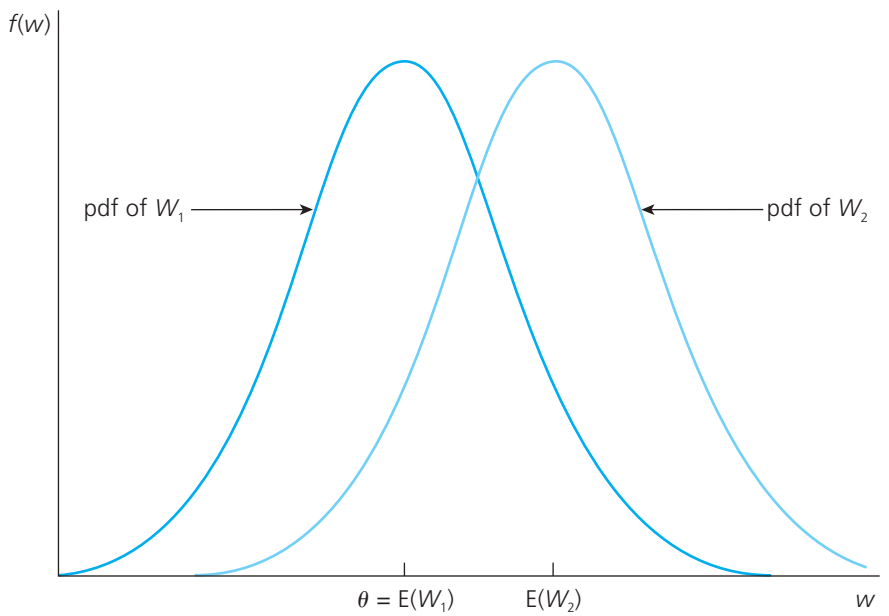
Figure C.1 shows two estimators; the first one is unbiased, and the second one has a positive bias.

The unbiasedness of an estimator and the size of any possible bias depend on the distribution of Y and on the function h. The distribution of Y is usually beyond our control (although we often choose a *model* for this distribution): it may be determined by nature or social forces. But the choice of the rule h is ours, and if we want an unbiased estimator, then we must choose h accordingly.

Some estimators can be shown to be unbiased quite generally. We now show that the sample average $\overline{Y}$ is an unbiased estimator of the population mean μ, regardless of the underlying population distribution. We use the properties of expected values (E.1 and E.2) that we covered in Section B-3:

$$\begin{aligned}\text{E}(\overline{Y}) &= \text{E}\left((1/n)\sum_{i=1}^{n} Y_i\right) = (1/n)\text{E}\left(\sum_{i=1}^{n} Y_i\right) = (1/n)\left(\sum_{i=1}^{n} \text{E}(Y_i)\right) \\ &= (1/n)\left(\sum_{i=1}^{n} \mu\right) = (1/n)(n\mu) = \mu.\end{aligned}$$

FIGURE C.1 An unbiased estimator, W_1, and an estimator with positive bias, W_2.



For hypothesis testing, we will need to estimate the variance σ^2 from a population with mean μ. Letting $\{Y_1, \ldots, Y_n\}$ denote the random sample from the population with $E(Y) = \mu$ and $Var(Y) = \sigma^2$, define the estimator as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2, \qquad \text{[C.5]}$$

which is usually called the **sample variance**. It can be shown that S^2 is unbiased for σ^2: $E(S^2) = \sigma^2$. The division by $n - 1$, rather than n, accounts for the fact that the mean μ is estimated rather than known. If μ were known, an unbiased estimator of σ^2 would be $n^{-1}\sum_{i=1}^{n}(Y_i - \mu)^2$, but μ is rarely known in practice.

Although unbiasedness has a certain appeal as a property for an estimator—indeed, its antonym, "biased," has decidedly negative connotations—it is not without its problems. One weakness of unbiasedness is that some reasonable, and even some very good, estimators are not unbiased. We will see an example shortly.

Another important weakness of unbiasedness is that unbiased estimators exist that are actually quite poor estimators. Consider estimating the mean μ from a population. Rather than using the sample average $\overline{Y}$ to estimate μ, suppose that, after collecting a sample of size n, we discard all of the observations except the first. That is, our estimator of μ is simply $W \equiv Y_1$. This estimator is unbiased because $E(Y_1) = \mu$. Hopefully, you sense that ignoring all but the first observation is not a prudent approach to estimation: it throws out most of the information in the sample. For example, with $n = 100$, we obtain 100 outcomes of the random variable Y, but then we use only the first of these to estimate $E(Y)$.

C-2c The Sampling Variance of Estimators

The example at the end of the previous subsection shows that we need additional criteria to evaluate estimators. Unbiasedness only ensures that the sampling distribution of an estimator has a mean value equal to the parameter it is supposed to be estimating. This is fine, but we also need to know how spread out the distribution of an estimator is. An estimator can be equal to θ, on average, but it can also be very far away with large probability. In Figure C.2, W_1 and W_2 are both unbiased estimators of θ. But the distribution of W_1 is more tightly centered about θ: the probability that W_1 is greater than any given distance from θ is less than the probability that W_2 is greater than that same distance from θ. Using W_1 as our estimator means that it is less likely that we will obtain a random sample that yields an estimate very far from θ.

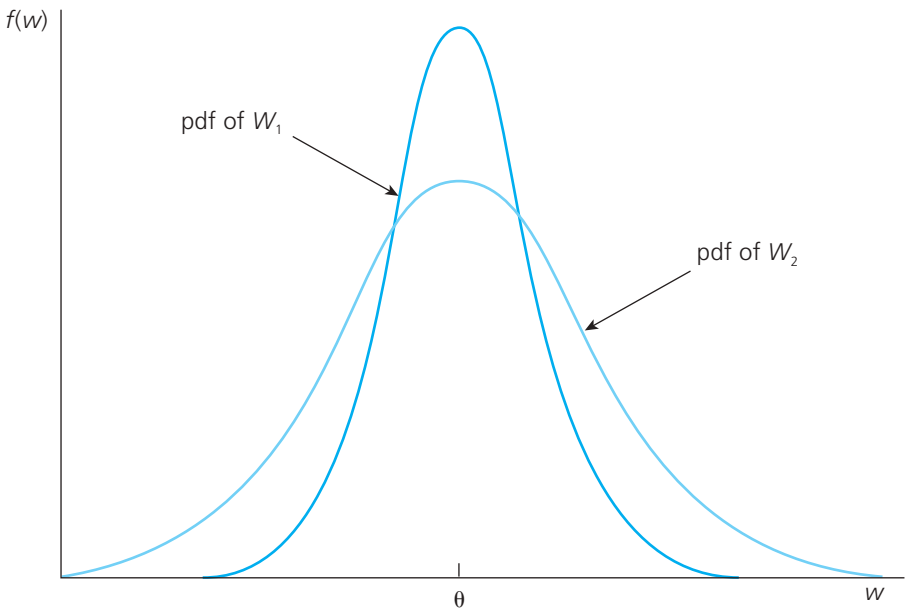
To summarize the situation shown in Figure C.2, we rely on the variance (or standard deviation) of an estimator. Recall that this gives a single measure of the dispersion in the distribution. The variance of an estimator is often called its **sampling variance** because it is the variance associated with a sampling distribution. Remember, the sampling variance is not a random variable; it is a constant, but it might be unknown.

We now obtain the variance of the sample average for estimating the mean μ from a population:

$$\begin{aligned} Var(\overline{Y}) &= Var\left((1/n)\sum_{i=1}^{n}Y_i\right) = (1/n^2)Var\left(\sum_{i=1}^{n}Y_i\right) = (1/n^2)\left(\sum_{i=1}^{n}Var(Y_i)\right) \\ &= (1/n^2)\left(\sum_{i=1}^{n}\sigma^2\right) = (1/n^2)(n\sigma^2) = \sigma^2/n. \end{aligned} \qquad \text{[C.6]}$$

Notice how we used the properties of variance from Sections B-3 and B-4 (VAR.2 and VAR.4), as well as the independence of the Y_i. To summarize: If $\{Y_i: i = 1, 2, \ldots, n\}$ is a random sample from a population with mean μ and variance σ^2, then $\overline{Y}$ has the same mean as the population, but its sampling variance equals the population variance, σ^2, divided by the sample size.

FIGURE C.2 The sampling distributions of two unbiased estimators of θ.



An important implication of $\text{Var}(\bar{Y}) = \sigma^2/n$ is that it can be made very close to zero by increasing the sample size n. This is a key feature of a reasonable estimator, and we return to it in Section C-3.

As suggested by Figure C.2, among unbiased estimators, we prefer the estimator with the smallest variance. This allows us to eliminate certain estimators from consideration. For a random sample from a population with mean μ and variance σ^2, we know that $\bar{Y}$ is unbiased and $\text{Var}(\bar{Y}) = \sigma^2/n$. What about the estimator Y_1, which is just the first observation drawn? Because Y_1 is a random draw from the population, $\text{Var}(Y_1) = \sigma^2$. Thus, the difference between $\text{Var}(Y_1)$ and $\text{Var}(\bar{Y})$ can be large even for small sample sizes. If $n = 10$, then $\text{Var}(Y_1)$ is 10 times as large as $\text{Var}(\bar{Y}) = \sigma^2/10$. This gives us a formal way of excluding Y_1 as an estimator of μ.

To emphasize this point, Table C.1 contains the outcome of a small simulation study. Using the statistical package Stata®, 20 random samples of size 10 were generated from a normal distribution, with $\mu = 2$ and $\sigma^2 = 1$; we are interested in estimating μ here. For each of the 20 random samples, we compute two estimates, y_1 and $\bar{y}$; these values are listed in Table C.1. As can be seen from the table, the values for y_1 are much more spread out than those for $\bar{y}$: y_1 ranges from -0.64 to 4.27, while $\bar{y}$ ranges only from 1.16 to 2.58. Further, in 16 out of 20 cases, $\bar{y}$ is closer than y_1 to $\mu = 2$. The average of y_1 across the simulations is about 1.89, while that for $\bar{y}$ is 1.96. The fact that these averages are close to 2 illustrates the unbiasedness of both estimators (and we could get these averages closer to 2 by doing more than 20 replications). But comparing just the average outcomes across random draws masks the fact that the sample average $\bar{Y}$ is far superior to Y_1 as an estimator of μ.

C-2d Efficiency

Comparing the variances of $\bar{Y}$ and Y_1 in the previous subsection is an example of a general approach to comparing different unbiased estimators.

Relative Efficiency. If W_1 and W_2 are two unbiased estimators of θ, W_1 is efficient relative to W_2 when $\text{Var}(W_1) \leq \text{Var}(W_2)$ for all θ, with strict inequality for at least one value of θ.

TABLE C.1 Simulation of Estimators for a Normal(μ, 1) Distribution with $\mu = 2$

Replication	y_1	$\bar{y}$
1	−0.64	1.98
2	1.06	1.43
3	4.27	1.65
4	1.03	1.88
5	3.16	2.34
6	2.77	2.58
7	1.68	1.58
8	2.98	2.23
9	2.25	1.96
10	2.04	2.11
11	0.95	2.15
12	1.36	1.93
13	2.62	2.02
14	2.97	2.10
15	1.93	2.18
16	1.14	2.10
17	2.08	1.94
18	1.52	2.21
19	1.33	1.16
20	1.21	1.75

Earlier, we showed that, for estimating the population mean μ, $\text{Var}(\bar{Y}) < \text{Var}(Y_1)$ for any value of σ^2 whenever $n > 1$. Thus, $\bar{Y}$ is efficient relative to Y_1 for estimating μ. We cannot always choose between unbiased estimators based on the smallest variance criterion: given two unbiased estimators of θ, one can have smaller variance from some values of θ, while the other can have smaller variance for other values of θ.

If we restrict our attention to a certain class of estimators, we can show that the sample average has the smallest variance. Problem C.2 asks you to show that $\bar{Y}$ has the smallest variance among all unbiased estimators that are also linear functions of $Y_1, Y_2, \ldots, Y_n$. The assumptions are that the Y_i have common mean and variance, and that they are pairwise uncorrelated.

If we do not restrict our attention to unbiased estimators, then comparing variances is meaningless. For example, when estimating the population mean μ, we can use a trivial estimator that is equal to zero, regardless of the sample that we draw. Naturally, the variance of this estimator is zero (because it is the same value for every random sample). But the bias of this estimator is $-\mu$, so it is a very poor estimator when $|\mu|$ is large.

One way to compare estimators that are not necessarily unbiased is to compute the **mean squared error (MSE)** of the estimators. If W is an estimator of θ, then the MSE of W is defined as $\text{MSE}(W) = \text{E}[(W - \theta)^2]$. The MSE measures how far, on average, the estimator is away from θ. It can be shown that $\text{MSE}(W) = \text{Var}(W) + [\text{Bias}(W)]^2$, so that MSE($W$) depends on the variance and bias (if any is present). This allows us to compare two estimators when one or both are biased.

C-3 Asymptotic or Large Sample Properties of Estimators

In Section C-2, we encountered the estimator Y_1 for the population mean μ, and we saw that, even though it is unbiased, it is a poor estimator because its variance can be much larger than that of the sample mean. One notable feature of Y_1 is that it has the same variance for any sample size. It seems reasonable to require any estimation procedure to improve as the sample size increases. For estimating a population mean μ, $\overline{Y}$ improves in the sense that its variance gets smaller as n gets larger; Y_1 does not improve in this sense.

We can rule out certain silly estimators by studying the *asymptotic* or *large sample* properties of estimators. In addition, we can say something positive about estimators that are not unbiased and whose variances are not easily found.

Asymptotic analysis involves approximating the features of the sampling distribution of an estimator. These approximations depend on the size of the sample. Unfortunately, we are necessarily limited in what we can say about how "large" a sample size is needed for asymptotic analysis to be appropriate; this depends on the underlying population distribution. But large sample approximations have been known to work well for sample sizes as small as $n = 20$.

C-3a Consistency

The first asymptotic property of estimators concerns how far the estimator is likely to be from the parameter it is supposed to be estimating as we let the sample size increase indefinitely.

Consistency. Let W_n be an estimator of θ based on a sample $Y_1, Y_2, \ldots, Y_n$ of size n. Then, W_n is a **consistent estimator** of θ if for every $\varepsilon > 0$,

$$P(|W_n - \theta| > \varepsilon) \to 0 \text{ as } n \to \infty. \qquad \text{[C.7]}$$

If W_n is not consistent for θ, then we say it is **inconsistent**.

When W_n is consistent, we also say that θ is the **probability limit** of W_n, written as $\text{plim}(W_n) = \theta$.

Unlike unbiasedness—which is a feature of an estimator for a given sample size—consistency involves the behavior of the sampling distribution of the estimator as the sample size n gets large. To emphasize this, we have indexed the estimator by the sample size in stating this definition, and we will continue with this convention throughout this section.

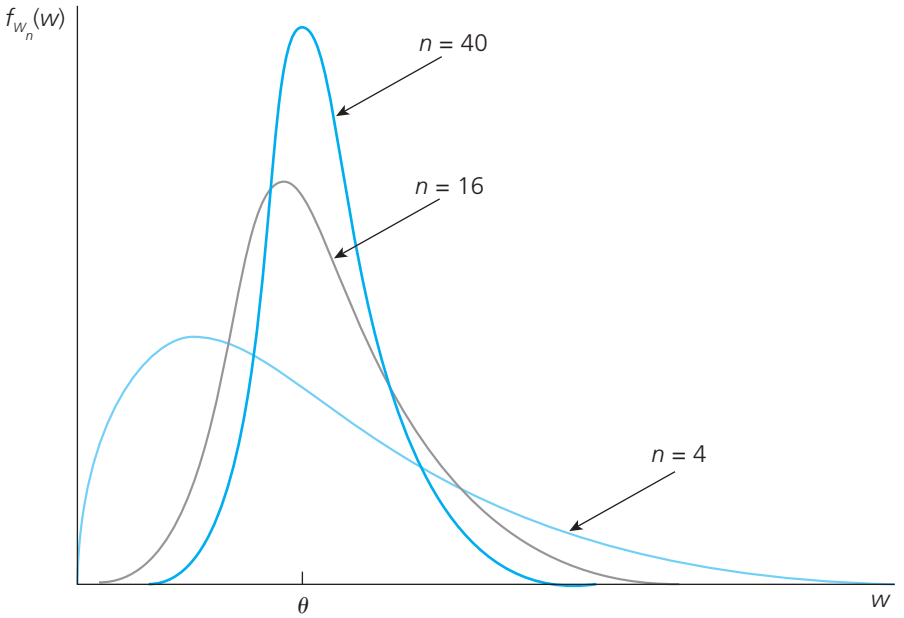
Equation (C.7) looks technical, and it can be rather difficult to establish based on fundamental probability principles. By contrast, interpreting (C.7) is straightforward. It means that the distribution of W_n becomes more and more concentrated about θ, which roughly means that for larger sample sizes, W_n is less and less likely to be very far from θ. This tendency is illustrated in Figure C.3.

If an estimator is not consistent, then it does not help us to learn about θ, even with an unlimited amount of data. For this reason, consistency is a minimal requirement of an estimator used in statistics or econometrics. We will encounter estimators that are consistent under certain assumptions and inconsistent when those assumptions fail. When estimators are inconsistent, we can usually find their probability limits, and it will be important to know how far these probability limits are from θ.

As we noted earlier, unbiased estimators are not necessarily consistent, but those whose variances shrink to zero as the sample size grows *are* consistent. This can be stated formally: If W_n is an unbiased estimator of θ and $\text{Var}(W_n) \to 0$ as $n \to \infty$, then $\text{plim}(W_n) = \theta$. Unbiased estimators that use the entire data sample will usually have a variance that shrinks to zero as the sample size grows, thereby being consistent.

A good example of a consistent estimator is the average of a random sample drawn from a population with mean μ and variance σ^2. We have already shown that the sample average is unbiased for μ.

FIGURE C.3 The sampling distributions of a consistent estimator for three sample sizes.



In Equation (C.6), we derived $\text{Var}(\bar{Y}_n) = \sigma^2/n$ for any sample size n. Therefore, $\text{Var}(\bar{Y}_n) \to 0$ as $n \to \infty$, so $\bar{Y}_n$ is a consistent estimator of μ (in addition to being unbiased).

The conclusion that $\bar{Y}_n$ is consistent for μ holds even if $\text{Var}(\bar{Y}_n)$ does not exist. This classic result is known as the **law of large numbers (LLN)**.

Law of Large Numbers. Let $Y_1, Y_2, \ldots, Y_n$ be independent, identically distributed random variables with mean μ. Then,

$$\text{plim}(\bar{Y}_n) = \mu. \qquad \text{[C.8]}$$

The law of large numbers means that, if we are interested in estimating the population average μ, we can get arbitrarily close to μ by choosing a sufficiently large sample. This fundamental result can be combined with basic properties of plims to show that fairly complicated estimators are consistent.

Property PLIM.1: Let θ be a parameter and define a new parameter, $\gamma = g(\theta)$, for some *continuous function* $g(\theta)$. Suppose that $\text{plim}(W_n) = \theta$. Define an estimator of γ by $G_n = g(W_n)$. Then,

$$\text{plim}(G_n) = \gamma. \qquad \text{[C.9]}$$

This is often stated as

$$\text{plim}\, g(W_n) = g(\text{plim}\, W_n) \qquad \text{[C.10]}$$

for a continuous function $g(\theta)$.

The assumption that $g(\theta)$ is continuous is a technical requirement that has often been described nontechnically as "a function that can be graphed without lifting your pencil from the paper." Because all the functions we encounter in this text are continuous, we do not provide a formal definition of a continuous function. Examples of continuous functions are $g(\theta) = a + b\theta$ for constants a and b, $g(\theta) = \theta^2$, $g(\theta) = 1/\theta$, $g(\theta) = \sqrt{\theta}$, $g(\theta) = \exp(\theta)$, and many variants on these. We will not need to mention the continuity assumption again.

As an important example of a consistent but biased estimator, consider estimating the standard deviation, σ, from a population with mean μ and variance σ^2. We already claimed that the sample variance $S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$ is unbiased for σ^2. Using the law of large numbers and some algebra, S_n^2 can also be shown to be consistent for σ^2. The natural estimator of $\sigma = \sqrt{\sigma^2}$ is $S_n = \sqrt{S_n^2}$ (where the square root is always the positive square root). S_n, which is called the **sample standard deviation**, is *not* an unbiased estimator because the expected value of the square root is *not* the square root of the expected value (see Section B-3). Nevertheless, by PLIM.1, $\text{plim}\, S_n = \sqrt{\text{plim}\, S_n^2} = \sqrt{\sigma^2} = \sigma$, so S_n is a consistent estimator of σ.

Here are some other useful properties of the probability limit:

Property PLIM.2: If $\text{plim}(T_n) = \alpha$ and $\text{plim}(U_n) = \beta$, then

(i) $\text{plim}(T_n + U_n) = \alpha + \beta$;
(ii) $\text{plim}(T_n U_n) = \alpha\beta$;
(iii) $\text{plim}(T_n/U_n) = \alpha/\beta$, provided $\beta \neq 0$.

These three facts about probability limits allow us to combine consistent estimators in a variety of ways to get other consistent estimators. For example, let $\{Y_1, \ldots, Y_n\}$ be a random sample of size n on annual earnings from the population of workers with a high school education and denote the population mean by μ_Y. Let $\{Z_1, \ldots, Z_n\}$ be a random sample on annual earnings from the population of workers with a college education and denote the population mean by μ_Z. We wish to estimate the percentage difference in annual earnings between the two groups, which is $\gamma = 100\cdot(\mu_Z - \mu_Y)/\mu_Y$. (This is the percentage by which average earnings for college graduates differs from average earnings for high school graduates.) Because $\bar{Y}_n$ is consistent for μ_Y and $\bar{Z}_n$ is consistent for μ_Z, it follows from PLIM.1 and part (iii) of PLIM.2 that

$$G_n \equiv 100\cdot(\bar{Z}_n - \bar{Y}_n)/\bar{Y}_n$$

is a consistent estimator of γ. G_n is just the percentage difference between $\bar{Z}_n$ and $\bar{Y}_n$ in the sample, so it is a natural estimator. G_n is not an unbiased estimator of γ, but it is still a good estimator except possibly when n is small.

C-3b Asymptotic Normality

Consistency is a property of point estimators. Although it does tell us that the distribution of the estimator is collapsing around the parameter as the sample size gets large, it tells us essentially nothing about the *shape* of that distribution for a given sample size. For constructing interval estimators and testing hypotheses, we need a way to approximate the distribution of our estimators. Most econometric estimators have distributions that are well approximated by a normal distribution for large samples, which motivates the following definition.

Asymptotic Normality. Let $\{Z_n: n = 1, 2, \ldots\}$ be a sequence of random variables, such that for all numbers z,

$$P(Z_n \leq z) \to \Phi(z) \text{ as } n \to \infty, \qquad \text{[C.11]}$$

where $\Phi(z)$ is the standard normal cumulative distribution function. Then, Z_n is said to have an *asymptotic standard normal distribution*. In this case, we often write $Z_n \overset{a}{\sim} \text{Normal}(0, 1)$. (The "$a$" above the tilde stands for "asymptotically" or "approximately.")

Property (C.11) means that the cumulative distribution function for Z_n gets closer and closer to the cdf of the standard normal distribution as the sample size n gets large. When **asymptotic normality** holds, for large n we have the approximation $P(Z_n \leq z) \approx \Phi(z)$. Thus, probabilities concerning Z_n can be approximated by standard normal probabilities.

The **central limit theorem (CLT)** is one of the most powerful results in probability and statistics. It states that the average from a random sample for *any* population (with finite variance), when standardized, has an asymptotic standard normal distribution.

Central Limit Theorem. Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample with mean μ and variance σ^2. Then,

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{Y}_n - \mu)}{\sigma} \qquad [C.12]$$

has an asymptotic standard normal distribution.

The variable Z_n in (C.12) is the standardized version of $\bar{Y}_n$: we have subtracted off $\mathrm{E}(\bar{Y}_n) = \mu$ and divided by $\mathrm{sd}(\bar{Y}_n) = \sigma/\sqrt{n}$. Thus, regardless of the population distribution of Y, Z_n has mean zero and variance one, which coincides with the mean and variance of the standard normal distribution. Remarkably, the entire distribution of Z_n gets arbitrarily close to the standard normal distribution as n gets large.

The second equality in equation (C.12) expresses the standardized variable as $\sqrt{n}(\bar{Y}_n - \mu)/\sigma$, which shows that we must multiply the difference between the sample mean and the population mean by the square root of the sample size in order to obtain a useful limiting distribution. Without the multiplication by $\sqrt{n}$, we would just have $(\bar{Y}_n - \mu)/\sigma$, which converges in probability to zero. In other words, the distribution of $(\bar{Y}_n - \mu)/\sigma$ simply collapses to a single point as $n \to \infty$, which we know cannot be a good approximation to the distribution of $(\bar{Y}_n - \mu)/\sigma$ for reasonable sample sizes. Multiplying by $\sqrt{n}$ ensures that the variance of Z_n remains constant. Practically, we often treat $\bar{Y}_n$ as being approximately normally distributed with mean μ and variance σ^2/n, and this gives us the correct statistical procedures because it leads to the standardized variable in equation (C.12).

Most estimators encountered in statistics and econometrics can be written as functions of sample averages, in which case we can apply the law of large numbers and the central limit theorem. When two consistent estimators have asymptotic normal distributions, we choose the estimator with the smallest asymptotic variance.

In addition to the standardized sample average in (C.12), many other statistics that depend on sample averages turn out to be asymptotically normal. An important one is obtained by replacing σ with its consistent estimator S_n in equation (C.12):

$$\frac{\bar{Y}_n - \mu}{S_n/\sqrt{n}} \qquad [C.13]$$

also has an approximate standard normal distribution for large n. The exact (finite sample) distributions of (C.12) and (C.13) are definitely not the same, but the difference is often small enough to be ignored for large n.

Throughout this section, each estimator has been subscripted by n to emphasize the nature of asymptotic or large sample analysis. Continuing this convention clutters the notation without providing additional insight, once the fundamentals of asymptotic analysis are understood. Henceforth, we drop the n subscript and rely on you to remember that estimators depend on the sample size, and properties such as consistency and asymptotic normality refer to the growth of the sample size without bound.

C-4 General Approaches to Parameter Estimation

Until this point, we have used the sample average to illustrate the finite and large sample properties of estimators. It is natural to ask: Are there general approaches to estimation that produce estimators with good properties, such as unbiasedness, consistency, and efficiency?

The answer is yes. A detailed treatment of various approaches to estimation is beyond the scope of this text; here, we provide only an informal discussion. A thorough discussion is given in Larsen and Marx (1986, Chapter 5).

C-4a Method of Moments

Given a parameter θ appearing in a population distribution, there are usually many ways to obtain unbiased and consistent estimators of θ. Trying all different possibilities and comparing them on the basis of the criteria in Sections C-2 and C-3 is not practical. Fortunately, some methods have been shown to have good general properties, and, for the most part, the logic behind them is intuitively appealing.

In the previous sections, we have studied the sample average as an unbiased estimator of the population average and the sample variance as an unbiased estimator of the population variance. These estimators are examples of **method of moments** estimators. Generally, method of moments estimation proceeds as follows. The parameter θ is shown to be related to some expected value in the distribution of Y, usually $E(Y)$ or $E(Y^2)$ (although more exotic choices are sometimes used). Suppose, for example, that the parameter of interest, θ, is related to the population mean as $\theta = g(\mu)$ for some function g. Because the sample average $\bar{Y}$ is an unbiased and consistent estimator of μ, it is natural to replace μ with $\bar{Y}$, which gives us the estimator $g(\bar{Y})$ of θ. The estimator $g(\bar{Y})$ is consistent for θ, and if $g(\mu)$ is a linear function of μ, then $g(\bar{Y})$ is unbiased as well. What we have done is replace the population moment, μ, with its sample counterpart, $\bar{Y}$. This is where the name "method of moments" comes from.

We cover two additional method of moments estimators that will be useful for our discussion of regression analysis. Recall that the covariance between two random variables X and Y is defined as $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$. The method of moments suggests estimating σ_{XY} by $n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$. This is a consistent estimator of σ_{XY}, but it turns out to be biased for essentially the same reason that the sample variance is biased if n, rather than $n - 1$, is used as the divisor. The **sample covariance** is defined as

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}). \qquad \text{[C.14]}$$

It can be shown that this is an unbiased estimator of σ_{XY}. (Replacing n with $n - 1$ makes no difference as the sample size grows indefinitely, so this estimator is still consistent.)

As we discussed in Section B-4, the covariance between two variables is often difficult to interpret. Usually, we are more interested in correlation. Because the population correlation is $\rho_{XY} = \sigma_{XY}/(\sigma_X\sigma_Y)$, the method of moments suggests estimating ρ_{XY} as

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)^{1/2}}, \qquad \text{[C.15]}$$

which is called the **sample correlation coefficient** (or *sample correlation* for short). Notice that we have canceled the division by $n - 1$ in the sample covariance and the sample standard deviations. In fact, we could divide each of these by n, and we would arrive at the same final formula.

It can be shown that the sample correlation coefficient is always in the interval $[-1,1]$, as it should be. Because S_{XY}, S_X, and S_Y are consistent for the corresponding population parameter, R_{XY} is a consistent estimator of the population correlation, ρ_{XY}. However, R_{XY} is a biased estimator for two reasons. First, S_X and S_Y are biased estimators of σ_X and σ_Y, respectively. Second, R_{XY} is a ratio of estimators, so it would not be unbiased, even if S_X and S_Y were. For our purposes, this is not important, although the fact that no unbiased estimator of ρ_{XY} exists is a classical result in mathematical statistics.

C-4b Maximum Likelihood

Another general approach to estimation is the method of *maximum likelihood*, a topic covered in many introductory statistics courses. A brief summary in the simplest case will suffice here. Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample from the population distribution $f(y; \theta)$. Because of the random

sampling assumption, the joint distribution of $\{Y_1, Y_2, \ldots, Y_n\}$ is simply the product of the densities: $f(y_1; \theta)f(y_2; \theta) \cdots f(y_n; \theta)$. In the discrete case, this is $P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$. Now, define the *likelihood function* as

$$L(\theta; Y_1, Y_2, \ldots, Y_n) = f(Y_1; \theta)f(Y_2; \theta) \cdot \cdots \cdot f(Y_n; \theta),$$

which is a random variable because it depends on the outcome of the random sample $\{Y_1, Y_2, \ldots, Y_n\}$. The **maximum likelihood estimator** of θ, call it W, is the value of θ that maximizes the likelihood function. (This is why we write L as a function of θ, followed by the random sample.) Clearly, this value depends on the random sample. The maximum likelihood principle says that, out of all the possible values for θ, the value that makes the likelihood of the observed data largest should be chosen. Intuitively, this is a reasonable approach to estimating θ.

Usually, it is more convenient to work with the **log-likelihood function**, which is obtained by taking the natural log of the likelihood function:

$$\mathcal{L}(\theta) = \log[L(\theta; Y_1, Y_2, \ldots, Y_n)] = \sum_{i=1}^{n} \log[f(Y_i; \theta)] = \sum_{i=1}^{n} \ell(\theta; X_i), \qquad \text{[C.16]}$$

where we use the fact that the log of the product is the sum of the logs. The function $\ell(\theta; X_i) = \log[f(Y_i; \theta)]$ is the log-likelihood function for random draw i. Because (C.16) is the sum of independent, identically distributed random variables, analyzing estimators that come from (C.16) is relatively easy.

Maximum likelihood estimation (MLE) is usually consistent and sometimes unbiased. But so are many other estimators. The widespread appeal of MLE is that it is generally the most asymptotically efficient estimator when the population model $f(y; \theta)$ is correctly specified. In addition, the MLE is sometimes the **minimum variance unbiased estimator**; that is, it has the smallest variance among all unbiased estimators of θ. [See Larsen and Marx (1986, Chapter 5) for verification of these claims.]

In Chapter 17, we will need maximum likelihood to estimate the parameters of more advanced econometric models. In econometrics, we are almost always interested in the distribution of Y conditional on a set of explanatory variables, say, $X_1, X_2, \ldots, X_k$. Then, we replace the density in (C.16) with $f(Y_i|X_{i1}, \ldots, X_{ik}; \theta_1, \ldots, \theta_p)$, where this density is allowed to depend on p parameters, $\theta_1, \ldots, \theta_p$. Fortunately, for successful application of maximum likelihood methods, we do not need to delve much into the computational issues or the large-sample statistical theory. Wooldridge (2010, Chapter 13) covers the theory of MLE.

C-4c Least Squares

A third kind of estimator, and one that plays a major role throughout the text, is called a **least squares estimator**. We have already seen an example of least squares: the sample mean, $\overline{Y}$, is a least squares estimator of the population mean, μ. We already know $\overline{Y}$ is a method of moments estimator. What makes it a least squares estimator? It can be shown that the value of m that makes the sum of squared deviations

$$\sum_{i=1}^{n} (Y_i - m)^2$$

as small as possible is $m = \overline{Y}$. Showing this is not difficult, but we omit the algebra.

For some important distributions, including the normal and the Bernoulli, the sample average $\overline{Y}$ is also the maximum likelihood estimator of the population mean μ. Thus, the principles of least squares, method of moments, and maximum likelihood often result in the *same* estimator. In other cases, the estimators are similar but not identical.

C-5 Interval Estimation and Confidence Intervals

C-5a The Nature of Interval Estimation

A point estimate obtained from a particular sample does not, by itself, provide enough information for testing economic theories or for informing policy discussions. A point estimate may be the researcher's best guess at the population value, but, by its nature, it provides no information about how close the estimate is "likely" to be to the population parameter. As an example, suppose a researcher reports, on the basis of a random sample of workers, that job training grants increase hourly wage by 6.4%. How are we to know whether or not this is close to the effect in the population of workers who could have been trained? Because we do not know the population value, we cannot know how close an estimate is for a particular sample. However, we can make statements involving probabilities, and this is where interval estimation comes in.

We already know one way of assessing the uncertainty in an estimator: find its sampling standard deviation. Reporting the standard deviation of the estimator, along with the point estimate, provides some information on the accuracy of our estimate. However, even if the problem of the standard deviation's dependence on unknown population parameters is ignored, reporting the standard deviation along with the point estimate makes no direct statement about where the population value is likely to lie in relation to the estimate. This limitation is overcome by constructing a **confidence interval**.

We illustrate the concept of a confidence interval with an example. Suppose the population has a Normal$(\mu, 1)$ distribution and let $\{Y_1, \ldots, Y_n\}$ be a random sample from this population. (We assume that the variance of the population is known and equal to unity for the sake of illustration; we then show what to do in the more realistic case that the variance is unknown.) The sample average, $\overline{Y}$, has a normal distribution with mean μ and variance $1/n$: $\overline{Y} \sim \text{Normal}(\mu, 1/n)$. From this, we can standardize $\overline{Y}$, and, because the standardized version of $\overline{Y}$ has a standard normal distribution, we have

$$\text{P}\left(-1.96 < \frac{\overline{Y} - \mu}{1/\sqrt{n}} < 1.96\right) = .95.$$

The event in parentheses is identical to the event $\overline{Y} - 1.96/\sqrt{n} < \mu < \overline{Y} + 1.96/\sqrt{n}$, so

$$\text{P}(\overline{Y} - 1.96/\sqrt{n} < \mu < \overline{Y} + 1.96\sqrt{n}) = .95. \qquad \text{[C.17]}$$

Equation (C.17) is interesting because it tells us that the probability that the random interval $[\overline{Y} - 1.96/\sqrt{n}, \overline{Y} + 1.96/\sqrt{n}]$ contains the population mean μ is .95, or 95%. This information allows us to construct an *interval estimate* of μ, which is obtained by plugging in the sample outcome of the average, $\bar{y}$. Thus,

$$[\bar{y} - 1.96/\sqrt{n}, \bar{y} + 1.96/\sqrt{n}] \qquad \text{[C.18]}$$

is an example of an interval estimate of μ. It is also called a 95% confidence interval. A shorthand notation for this interval is $\bar{y} \pm 1.96/\sqrt{n}$.

The confidence interval in equation (C.18) is easy to compute, once the sample data $\{y_1, y_2, \ldots, y_n\}$ are observed; $\bar{y}$ is the only factor that depends on the data. For example, suppose that $n = 16$ and the average of the 16 data points is 7.3. Then, the 95% confidence interval for μ is $7.3 \pm 1.96/\sqrt{16} = 7.3 \pm .49$, which we can write in interval form as [6.81,7.79]. By construction, $\bar{y} = 7.3$ is in the center of this interval.

Unlike its computation, the meaning of a confidence interval is more difficult to understand. When we say that equation (C.18) is a 95% confidence interval for μ, we mean that the *random* interval

$$[\overline{Y} - 1.96/\sqrt{n}, \overline{Y} + 1.96/\sqrt{n}] \qquad \text{[C.19]}$$

contains μ with probability .95. In other words, *before* the random sample is drawn, there is a 95% chance that (C.19) contains μ. Equation (C.19) is an example of an **interval estimator**. It is a random interval, because the endpoints change with different samples.

A confidence interval is often interpreted as follows: "The probability that μ is in the interval (C.18) is .95." This is incorrect. Once the sample has been observed and $\bar{y}$ has been computed, the limits of the confidence interval are simply numbers (6.81 and 7.79 in the example just given). The population parameter, μ, though unknown, is also just some number. Therefore, μ either is or is not in the interval (C.18) (and we will never know with certainty which is the case). Probability plays no role once the confidence interval is computed for the particular data at hand. The probabilistic interpretation comes from the fact that for 95% of all random samples, the constructed confidence interval will contain μ.

To emphasize the meaning of a confidence interval, Table C.2 contains calculations for 20 random samples (or replications) from the Normal(2,1) distribution with sample size $n = 10$. For each of the 20 samples, $\bar{y}$ is obtained, and (C.18) is computed as $\bar{y} \pm 1.96/\sqrt{10} = \bar{y} \pm .62$ (each rounded to two decimals). As you can see, the interval changes with each random sample. Nineteen of the 20 intervals contain the population value of μ. Only for replication number 19 is μ not in the confidence interval. In other words, 95% of the samples result in a confidence interval that contains μ. This did not have to be the case with only 20 replications, but it worked out that way for this particular simulation.

TABLE C.2 Simulated Confidence Intervals from a Normal(μ, 1) Distribution with $\mu = 2$

Replication	$\bar{y}$	95% Interval	Contains μ?
1	1.98	(1.36,2.60)	Yes
2	1.43	(0.81,2.05)	Yes
3	1.65	(1.03,2.27)	Yes
4	1.88	(1.26,2.50)	Yes
5	2.34	(1.72,2.96)	Yes
6	2.58	(1.96,3.20)	Yes
7	1.58	(.96,2.20)	Yes
8	2.23	(1.61,2.85)	Yes
9	1.96	(1.34,2.58)	Yes
10	2.11	(1.49,2.73)	Yes
11	2.15	(1.53,2.77)	Yes
12	1.93	(1.31,2.55)	Yes
13	2.02	(1.40,2.64)	Yes
14	2.10	(1.48,2.72)	Yes
15	2.18	(1.56,2.80)	Yes
16	2.10	(1.48,2.72)	Yes
17	1.94	(1.32,2.56)	Yes
18	2.21	(1.59,2.83)	Yes
19	1.16	(.54,1.78)	No
20	1.75	(1.13,2.37)	Yes

C-5b Confidence Intervals for the Mean from a Normally Distributed Population

The confidence interval derived in equation (C.18) helps illustrate how to construct and interpret confidence intervals. In practice, equation (C.18) is not very useful for the mean of a normal population because it assumes that the variance is known to be unity. It is easy to extend (C.18) to the case where the standard deviation σ is known to be any value: the 95% confidence interval is

$$[\bar{y} - 1.96\sigma/\sqrt{n}, \bar{y} + 1.96\sigma/\sqrt{n}]. \qquad [C.20]$$

Therefore, provided σ is known, a confidence interval for μ is readily constructed. To allow for unknown σ, we must use an estimate. Let

$$s = \left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{1/2} \qquad [C.21]$$

denote the sample standard deviation. Then, we obtain a confidence interval that depends entirely on the observed data by replacing σ in equation (C.20) with its estimate, s. Unfortunately, this does not preserve the 95% level of confidence because s depends on the particular sample. In other words, the random interval $[\bar{Y} \pm 1.96(S/\sqrt{n})]$ no longer contains μ with probability .95 because the constant σ has been replaced with the random variable S.

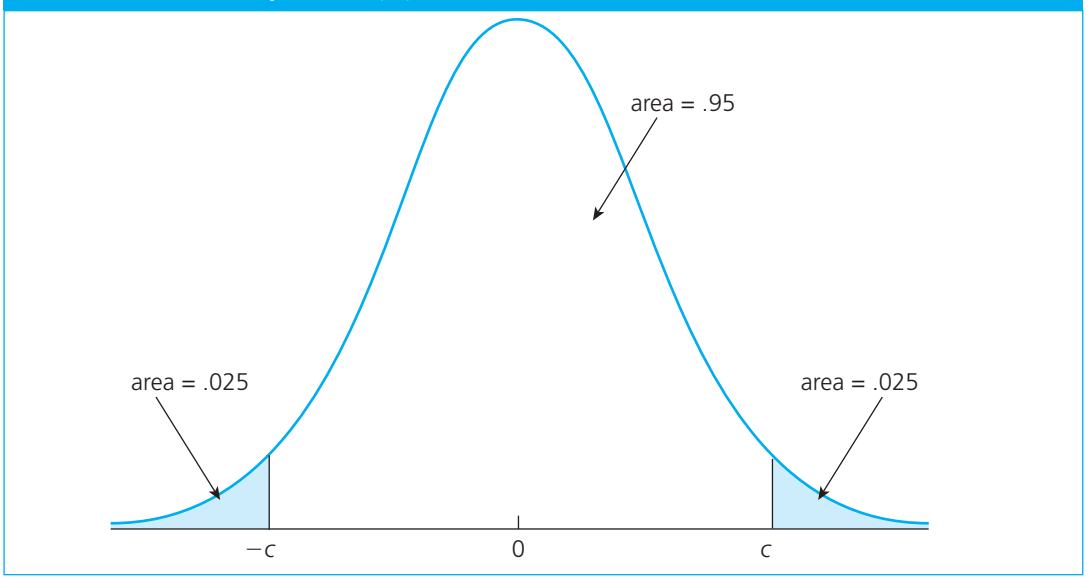
How should we proceed? Rather than using the standard normal distribution, we must rely on the t distribution. The t distribution arises from the fact that

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \qquad [C.22]$$

where $\bar{Y}$ is the sample average and S is the sample standard deviation of the random sample $\{Y_1, \ldots, Y_n\}$. We will not prove (C.22); a careful proof can be found in a variety of places [for example, Larsen and Marx (1986, Chapter 7)].

To construct a 95% confidence interval, let c denote the 97.5th percentile in the t_{n-1} distribution. In other words, c is the value such that 95% of the area in the t_{n-1} is between $-c$ and c: $P(-c < t_{n-1} < c) = .95$. (The value of c depends on the degrees of freedom $n - 1$, but we do not

FIGURE C.4 The 97.5th percentile, c, in a t distribution.



make this explicit.) The choice of c is illustrated in Figure C.4. Once c has been properly chosen, the random interval $[\bar{Y} - c \cdot S/\sqrt{n}, \bar{Y} + c \cdot S/\sqrt{n}]$ contains μ with probability .95. For a particular sample, the 95% confidence interval is calculated as

$$[\bar{y} - c \cdot s/\sqrt{n}, \bar{y} + c \cdot s/\sqrt{n}]. \qquad \text{[C.23]}$$

The values of c for various degrees of freedom can be obtained from Table G.2 in Statistical Tables. For example, if $n = 20$, so that the df is $n - 1 = 19$, then $c = 2.093$. Thus, the 95% confidence interval is $[\bar{y} \pm 2.093(s/\sqrt{20})]$, where $\bar{y}$ and s are the values obtained from the sample. Even if $s = \sigma$ (which is very unlikely), the confidence interval in (C.23) is wider than that in (C.20) because $c > 1.96$. For small degrees of freedom, (C.23) is much wider.

More generally, let c_α denote the $100(1 - \alpha)$ percentile in the t_{n-1} distribution. Then, a $100(1 - \alpha)\%$ confidence interval is obtained as

$$[\bar{y} - c_{\alpha/2} s/\sqrt{n}, \bar{y} + c_{\alpha/2} s/\sqrt{n}]. \qquad \text{[C.24]}$$

Obtaining $c_{\alpha/2}$ requires choosing α and knowing the degrees of freedom $n - 1$; then, Table G.2 can be used. For the most part, we will concentrate on 95% confidence intervals.

There is a simple way to remember how to construct a confidence interval for the mean of a normal distribution. Recall that $\text{sd}(\bar{Y}) = \sigma/\sqrt{n}$. Thus, $s/\sqrt{n}$ is the point estimate of $\text{sd}(\bar{Y})$. The associated random variable, $S/\sqrt{n}$, is sometimes called the **standard error** of $\bar{Y}$. Because what shows up in formulas is the point estimate $s/\sqrt{n}$, we define the standard error of $\bar{y}$ as $\text{se}(\bar{y}) = s/\sqrt{n}$. Then, (C.24) can be written in shorthand as

$$[\bar{y} \pm c_{\alpha/2} \cdot \text{se}(\bar{y})]. \qquad \text{[C.25]}$$

This equation shows why the notion of the standard error of an estimate plays an important role in econometrics.

EXAMPLE C.2 Effect of Job Training Grants on Worker Productivity

Holzer, Block, Cheatham, and Knott (1993) studied the effects of job training grants on worker productivity by collecting information on "scrap rates" for a sample of Michigan manufacturing firms receiving job training grants in 1988. Table C.3 lists the scrap rates—measured as number of items per 100 produced that are not usable and therefore need to be scrapped—for 20 firms. Each of these firms received a job training grant in 1988; there were no grants awarded in 1987. We are interested in constructing a confidence interval for the change in the scrap rate from 1987 to 1988 for the population of all manufacturing firms that could have received grants.

We assume that the change in scrap rates has a normal distribution. Because $n = 20$, a 95% confidence interval for the mean change in scrap rates μ is $[\bar{y} \pm 2.093 \cdot \text{se}(\bar{y})]$, where $\text{se}(\bar{y}) = s/\sqrt{n}$. The value 2.093 is the 97.5^{th} percentile in a t_{19} distribution. For the particular sample values, $\bar{y} = -1.15$ and $\text{se}(\bar{y}) = .54$ (each rounded to two decimals), so the 95% confidence interval is $[-2.28, -.02]$. The value zero is excluded from this interval, so we conclude that, with 95% confidence, the average change in scrap rates in the population is not zero.

TABLE C.3 Scrap Rates for 20 Michigan Manufacturing Firms

Firm	1987	1988	Change
1	10	3	−7
2	1	1	0
3	6	5	−1
4	.45	.5	.05
5	1.25	1.54	.29
6	1.3	1.5	.2
7	1.06	.8	−.26
8	3	2	−1
9	8.18	.67	−7.51
10	1.67	1.17	−.5
11	.98	.51	−.47
12	1	.5	−.5
13	.45	.61	.16
14	5.03	6.7	1.67
15	8	4	−4
16	9	7	−2
17	18	19	1
18	.28	.2	−.08
19	7	5	−2
20	3.97	3.83	−.14
Average	4.38	3.23	−1.15

At this point, Example C.2 is mostly illustrative because it has some potentially serious flaws as an econometric analysis. Most importantly, it assumes that any systematic reduction in scrap rates is due to the job training grants. But many things can happen over the course of the year to change worker productivity. From this analysis, we have no way of knowing whether the fall in average scrap rates is attributable to the job training grants or if, at least partly, some external force is responsible.

C-5c A Simple Rule of Thumb for a 95% Confidence Interval

The confidence interval in (C.25) can be computed for any sample size and any confidence level. As we saw in Section B-5, the t distribution approaches the standard normal distribution as the degrees of freedom gets large. In particular, for $\alpha = .05$, $c_{\alpha/2} \to 1.96$ as $n \to \infty$, although $c_{\alpha/2}$ is always greater than 1.96 for each n. A *rule of thumb* for an approximate 95% confidence interval is

$$[\bar{y} \pm 2 \cdot \text{se}(\bar{y})]. \qquad \text{[C.26]}$$

In other words, we obtain $\bar{y}$ and its standard error and then compute $\bar{y}$ plus or minus twice its standard error to obtain the confidence interval. This is slightly too wide for very large n, and it is too narrow for small n. As we can see from Example C.2, even for n as small as 20, (C.26) is in the ballpark for a 95% confidence interval for the mean from a normal distribution. This means we can get pretty close to a 95% confidence interval without having to refer to t tables.

C-5d Asymptotic Confidence Intervals for Nonnormal Populations

In some applications, the population is clearly nonnormal. A leading case is the Bernoulli distribution, where the random variable takes on only the values zero and one. In other cases, the nonnormal population has no standard distribution. This does not matter, provided the sample size is sufficiently large for the central limit theorem to give a good approximation for the distribution of the sample average $\bar{Y}$. For large n, an *approximate* 95% confidence interval is

$$[\bar{y} \pm 1.96 \cdot \text{se}(\bar{y})], \qquad \text{[C.27]}$$

where the value 1.96 is the 97.5$^{\text{th}}$ percentile in the standard normal distribution. Mechanically, computing an approximate confidence interval does not differ from the normal case. A slight difference is that the number multiplying the standard error comes from the standard normal distribution, rather than the t distribution, because we are using asymptotics. Because the t distribution approaches the standard normal as the df increases, equation (C.25) is also perfectly legitimate as an approximate 95% interval; some prefer this to (C.27) because the former is exact for normal populations.

EXAMPLE C.3 Race Discrimination in Hiring

The Urban Institute conducted a study in 1988 in Washington, D.C., to examine the extent of race discrimination in hiring. Five pairs of people interviewed for several jobs. In each pair, one person was black and the other person was white. They were given résumés indicating that they were virtually the same in terms of experience, education, and other factors that determine job qualification. The idea was to make individuals as similar as possible with the exception of race. Each person in a pair interviewed for the same job, and the researchers recorded which applicant received a job offer. This is an example of a *matched pairs analysis*, where each trial consists of data on two people (or two firms, two cities, and so on) that are thought to be similar in many respects but different in one important characteristic.

Let θ_B denote the probability that the black person is offered a job and let θ_W be the probability that the white person is offered a job. We are primarily interested in the difference, $\theta_B - \theta_W$. Let B_i denote a Bernoulli variable equal to one if the black person gets a job offer from employer i, and zero otherwise. Similarly, $W_i = 1$ if the white person gets a job offer from employer i, and zero otherwise. Pooling across the five pairs of people, there were a total of $n = 241$ trials (pairs of interviews with employers). Unbiased estimators of θ_B and θ_W are $\overline{B}$ and $\overline{W}$, the fractions of interviews for which blacks and whites were offered jobs, respectively.

To put this into the framework of computing a confidence interval for a population mean, define a new variable $Y_i = B_i - W_i$. Now, Y_i can take on three values: -1 if the black person did not get the job but the white person did, 0 if both people either did or did not get the job, and 1 if the black person got the job and the white person did not. Then, $\mu \equiv \text{E}(Y_i) = \text{E}(B_i) - \text{E}(W_i) = \theta_B - \theta_W$.

The distribution of Y_i is certainly not normal—it is discrete and takes on only three values. Nevertheless, an approximate confidence interval for $\theta_B - \theta_W$ can be obtained by using large sample methods.

The data from the Urban Institute audit study are in the file AUDIT. Using the 241 observed data points, $\overline{b} = .224$ and $\overline{w} = .357$, so $\bar{y} = .224 - .357 = -.133$. Thus, 22.4% of black applicants were offered jobs, while 35.7% of white applicants were offered jobs. This is *prima facie* evidence of discrimination against blacks, but we can learn much more by computing a confidence interval for μ. To compute an approximate 95% confidence interval, we need the sample standard deviation. This turns out to be $s = .482$ [using equation (C.21)]. Using (C.27), we obtain a 95% CI for $\mu = \theta_B - \theta_W$ as $-.133 \pm 1.96(.482/\sqrt{241}) = -.133 \pm .031 = [-.164, -.102]$. The approximate 99% CI is $-.133 \pm 2.58(.482/\sqrt{241}) = [-.213, -.053]$. Naturally, this contains a wider range of values than the 95% CI. But even the 99% CI does not contain the value zero. Thus, we are very confident that the population difference $\theta_B - \theta_W$ is not zero.

Before we turn to hypothesis testing, it is useful to review the various population and sample quantities that measure the spreads in the population distributions and the sampling distributions of the estimators. These quantities appear often in statistical analysis, and extensions of them are important for the regression analysis in the main text. The quantity σ is the (unknown) population standard deviation; it is a measure of the spread in the distribution of Y. When we divide σ by $\sqrt{n}$, we obtain the **sampling standard deviation** of $\overline{Y}$ (the sample average). While σ is a fixed feature of the population, $\text{sd}(\overline{Y}) = \sigma/\sqrt{n}$ shrinks to zero as $n \to \infty$: our estimator of μ gets more and more precise as the sample size grows.

The estimate of σ for a particular sample, s, is called the sample standard deviation because it is obtained from the sample. (We also call the underlying random variable, S, which changes across different samples, the sample standard deviation.) Like $\bar{y}$ as an estimate of μ, s is our "best guess" at σ given the sample at hand. The quantity $s/\sqrt{n}$ is what we call the standard error of $\bar{y}$, and it is our best estimate of $\sigma/\sqrt{n}$. Confidence intervals for the population parameter μ depend directly on $\text{se}(\bar{y}) = s/\sqrt{n}$. Because this standard error shrinks to zero as the sample size grows, a larger sample size generally means a smaller confidence interval. Thus, we see clearly that one benefit of more data is that they result in narrower confidence intervals. The notion of the standard error of an estimate, which in the vast majority of cases shrinks to zero at the rate $1/\sqrt{n}$, plays a fundamental role in hypothesis testing (as we will see in the next section) and for confidence intervals and testing in the context of multiple regression (as discussed in Chapter 4).

C-6 Hypothesis Testing

So far, we have reviewed how to evaluate point estimators, and we have seen—in the case of a population mean—how to construct and interpret confidence intervals. But sometimes the question we are interested in has a definite yes or no answer. Here are some examples: (1) Does a job training program effectively increase average worker productivity? (see Example C.2); (2) Are blacks discriminated against in hiring? (see Example C.3); (3) Do stiffer state drunk driving laws reduce the number of drunk driving arrests? Devising methods for answering such questions, using a sample of data, is known as hypothesis testing.

C-6a Fundamentals of Hypothesis Testing

To illustrate the issues involved with hypothesis testing, consider an election example. Suppose there are two candidates in an election, Candidates A and B. Candidate A is reported to have received 42% of the popular vote, while Candidate B received 58%. These are supposed to represent the true percentages in the voting population, and we treat them as such.

Candidate A is convinced that more people must have voted for him, so he would like to investigate whether the election was rigged. Knowing something about statistics, Candidate A hires a consulting agency to randomly sample 100 voters to record whether or not each person voted for him. Suppose that, for the sample collected, 53 people voted for Candidate A. This sample estimate of 53% clearly exceeds the reported population value of 42%. Should Candidate A conclude that the election was indeed a fraud?

While it appears that the votes for Candidate A were undercounted, we cannot be certain. Even if only 42% of the population voted for Candidate A, it is possible that, in a sample of 100, we observe 53 people who did vote for Candidate A. The question is: How *strong* is the sample evidence against the officially reported percentage of 42%?

One way to proceed is to set up a **hypothesis test**. Let θ denote the true proportion of the population voting for Candidate A. The hypothesis that the reported results are accurate can be stated as

$$H_0: \theta = .42 \qquad \text{[C.28]}$$

This is an example of a **null hypothesis**. We always denote the null hypothesis by H_0. In hypothesis testing, the null hypothesis plays a role similar to that of a defendant on trial in many judicial systems: just as a defendant is presumed to be innocent until proven guilty, the null hypothesis is presumed to be true until the data strongly suggest otherwise. In the current example, Candidate A must present fairly strong evidence against (C.28) in order to win a recount.

The **alternative hypothesis** in the election example is that the true proportion voting for Candidate A in the election is greater than .42:

$$H_1: \theta > .42. \qquad \text{[C.29]}$$

In order to conclude that H_0 is false and that H_1 is true, we must have evidence "beyond reasonable doubt" against H_0. How many votes out of 100 would be needed before we feel the evidence is strongly against H_0? Most would agree that observing 43 votes out of a sample of 100 is not enough to overturn the original election results; such an outcome is well within the expected sampling variation. On the other hand, we do not need to observe 100 votes for Candidate A to cast doubt on H_0. Whether 53 out of 100 is enough to reject H_0 is much less clear. The answer depends on how we quantify "beyond reasonable doubt."

Before we turn to the issue of quantifying uncertainty in hypothesis testing, we should head off some possible confusion. You may have noticed that the hypotheses in equations (C.28) and (C.29) do not exhaust all possibilities: it could be that θ is less than .42. For the application at hand, we are not particularly interested in that possibility; it has nothing to do with overturning the results of the election. Therefore, we can just state at the outset that we are ignoring alternatives θ with $\theta < .42$. Nevertheless, some authors prefer to state null and alternative hypotheses so that they are exhaustive, in which case our null hypothesis should be $H_0: \theta \leq .42$. Stated in this way, the null hypothesis is a *composite* null hypothesis because it allows for more than one value under H_0. [By contrast, equation (C.28) is an example of a *simple* null hypothesis.] For these kinds of examples, it does not matter whether we state the null as in (C.28) or as a composite null: the most difficult value to reject if $\theta \leq .42$ is $\theta = .42$. (That is, if we reject the value $\theta = .42$, against $\theta > .42$, then logically we must reject any value less than .42.) Therefore, our testing procedure based on (C.28) leads to the same test as if $H_0: \theta \leq .42$. In this text, we always state a null hypothesis as a simple null hypothesis.

In hypothesis testing, we can make two kinds of mistakes. First, we can reject the null hypothesis when it is in fact true. This is called a **Type I error**. In the election example, a Type I error occurs if we reject H_0 when the true proportion of people voting for Candidate A is in fact .42. The second kind of error is failing to reject H_0 when it is actually false. This is called a **Type II error**. In the election example, a Type II error occurs if $\theta > .42$ but we fail to reject H_0.

After we have made the decision of whether or not to reject the null hypothesis, we have either decided correctly or we have committed an error. We will never know with certainty whether an error was committed. However, we can compute the *probability* of making either a Type I or a Type II error. Hypothesis testing rules are constructed to make the probability of committing a Type I error fairly small. Generally, we define the **significance level** (or simply the *level*) of a test as the probability of a Type I error; it is typically denoted by α. Symbolically, we have

$$\alpha = P(\text{Reject } H_0 | H_0). \qquad \text{[C.30]}$$

The right-hand side is read as: "The probability of rejecting H_0 given that H_0 is true."

Classical hypothesis testing requires that we initially specify a significance level for a test. When we specify a value for α, we are essentially quantifying our tolerance for a Type I error. Common values for α are .10, .05, and .01. If $\alpha = .05$, then the researcher is willing to falsely reject H_0 5% of the time, in order to detect deviations from H_0.

Once we have chosen the significance level, we would then like to minimize the probability of a Type II error. Alternatively, we would like to maximize the **power of a test** against all relevant alternatives. The power of a test is just one minus the probability of a Type II error. Mathematically,

$$\pi(\theta) = P(\text{Reject } H_0 | \theta) = 1 - P(\text{Type II} | \theta),$$

where θ denotes the actual value of the parameter. Naturally, we would like the power to equal unity whenever the null hypothesis is false. But this is impossible to achieve while keeping the significance level small. Instead, we choose our tests to maximize the power for a given significance level.

C-6b Testing Hypotheses about the Mean in a Normal Population

In order to test a null hypothesis against an alternative, we need to choose a test statistic (or statistic, for short) and a critical value. The choices for the statistic and critical value are based on convenience and on the desire to maximize power given a significance level for the test. In this subsection, we review how to test hypotheses for the mean of a normal population.

A **test statistic**, denoted T, is some function of the random sample. When we compute the statistic for a particular outcome, we obtain an outcome of the test statistic, which we will denote by t.

Given a test statistic, we can define a rejection rule that determines when H_0 is rejected in favor of H_1. In this text, all rejection rules are based on comparing the value of a test statistic, t, to a **critical value**, c. The values of t that result in rejection of the null hypothesis are collectively known as the **rejection region**. To determine the critical value, we must first decide on a significance level of the test. Then, given α, the critical value associated with α is determined by the distribution of T, *assuming* that H_0 is true. We will write this critical value as c, suppressing the fact that it depends on α.

Testing hypotheses about the mean μ from a Normal(μ, σ^2) population is straightforward. The null hypothesis is stated as

$$H_0: \mu = \mu_0, \tag{C.31}$$

where μ_0 is a value that we specify. In the majority of applications, $\mu_0 = 0$, but the general case is no more difficult.

The rejection rule we choose depends on the nature of the alternative hypothesis. The three alternatives of interest are

$$H_1: \mu > \mu_0, \tag{C.32}$$

$$H_1: \mu < \mu_0, \tag{C.33}$$

and

$$H_1: \mu \neq \mu_0. \tag{C.34}$$

Equation (C.32) gives a **one-sided alternative**, as does (C.33). When the alternative hypothesis is (C.32), the null is effectively $H_0: \mu \leq \mu_0$, because we reject H_0 only when $\mu > \mu_0$. This is appropriate when we are interested in the value of μ only when μ is at least as large as μ_0. Equation (C.34) is a **two-sided alternative**. This is appropriate when we are interested in any departure from the null hypothesis.

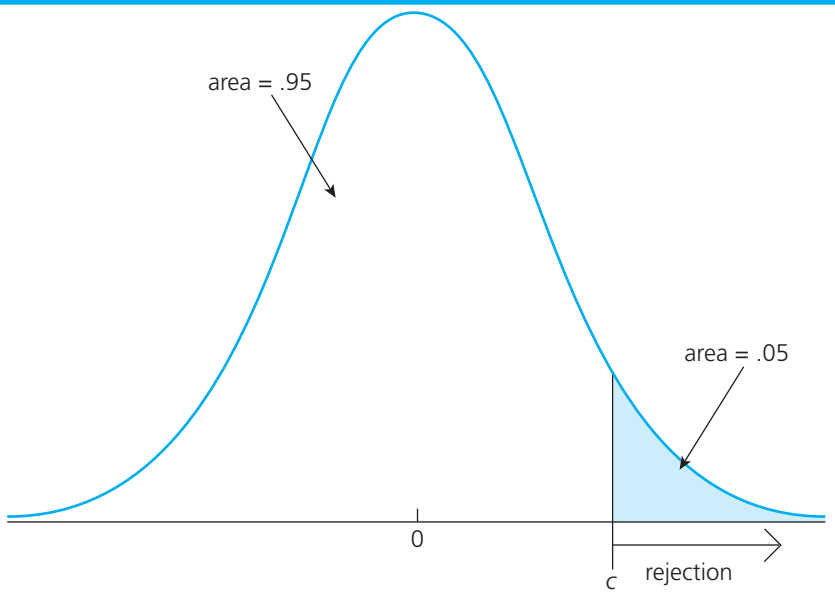
Consider first the alternative in (C.32). Intuitively, we should reject H_0 in favor of H_1 when the value of the sample average, $\bar{y}$, is "sufficiently" greater than μ_0. But how should we determine when $\bar{y}$ is large enough for H_0 to be rejected at the chosen significance level? This requires knowing the probability of rejecting the null hypothesis when it is true. Rather than working directly with $\bar{y}$, we use its standardized version, where σ is replaced with the sample standard deviation, s:

$$t = \sqrt{n}(\bar{y} - \mu_0)/s = (\bar{y} - \mu_0)/\text{se}(\bar{y}), \tag{C.35}$$

where $\text{se}(\bar{y}) = s/\sqrt{n}$ is the standard error of $\bar{y}$. Given the sample of data, it is easy to obtain t. We work with t because, under the null hypothesis, the random variable

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S$$

FIGURE C.5 Rejection region for a 5% significance level test against the one-sided alternative $\mu > \mu_0$.



has a t_{n-1} distribution. Now, suppose we have settled on a 5% significance level. Then, the critical value c is chosen so that $P(T > c|H_0) = .05$; that is, the probability of a Type I error is 5%. Once we have found c, the rejection rule is

$$t > c, \qquad [C.36]$$

where c is the $100(1 - \alpha)$ percentile in a t_{n-1} distribution; as a percent, the significance level is $100 \cdot \alpha\%$. This is an example of a **one-tailed test** because the rejection region is in one tail of the t distribution. For a 5% significance level, c is the 95th percentile in the t_{n-1} distribution; this is illustrated in Figure C.5. A different significance level leads to a different critical value.

The statistic in equation (C.35) is often called the ***t* statistic** for testing H_0: $\mu = \mu_0$. The t statistic measures the distance from $\bar{y}$ to μ_0 *relative* to the standard error of $\bar{y}$, $se(\bar{y})$.

EXAMPLE C.4 Effect of Enterprise Zones on Business Investments

In the population of cities granted enterprise zones in a particular state [see Papke (1994) for Indiana], let Y denote the percentage change in investment from the year before to the year after a city became an enterprise zone. Assume that Y has a Normal(μ, σ^2) distribution. The null hypothesis that enterprise zones have no effect on business investment is H_0: $\mu = 0$; the alternative that they have a positive effect is H_1: $\mu > 0$. (We assume that they do not have a negative effect.) Suppose that we wish to test H_0 at the 5% level. The test statistic in this case is

$$t = \frac{\bar{y}}{s/\sqrt{n}} = \frac{\bar{y}}{se(\bar{y})}. \qquad [C.37]$$

Suppose that we have a sample of 36 cities that are granted enterprise zones. Then, the critical value is $c = 1.69$ (see Table G.2), and we reject H_0 in favor of H_1 if $t > 1.69$. Suppose that the sample yields $\bar{y} = 8.2$ and $s = 23.9$. Then, $t \approx 2.06$, and H_0 is therefore rejected at the 5% level. Thus, we conclude

that, at the 5% significance level, enterprise zones have an effect on average investment. The 1% critical value is 2.44, so H_0 is not rejected at the 1% level. The same caveat holds here as in Example C.2: we have not controlled for other factors that might affect investment in cities over time, so we cannot claim that the effect is causal.

The rejection rule is similar for the one-sided alternative (C.33). A test with a significance level of $100 \cdot \alpha\%$ rejects H_0 against (C.33) whenever

$$t < -c; \qquad [C.38]$$

in other words, we are looking for negative values of the t statistic—which implies $\bar{y} < \mu_0$—that are sufficiently far from zero to reject H_0.

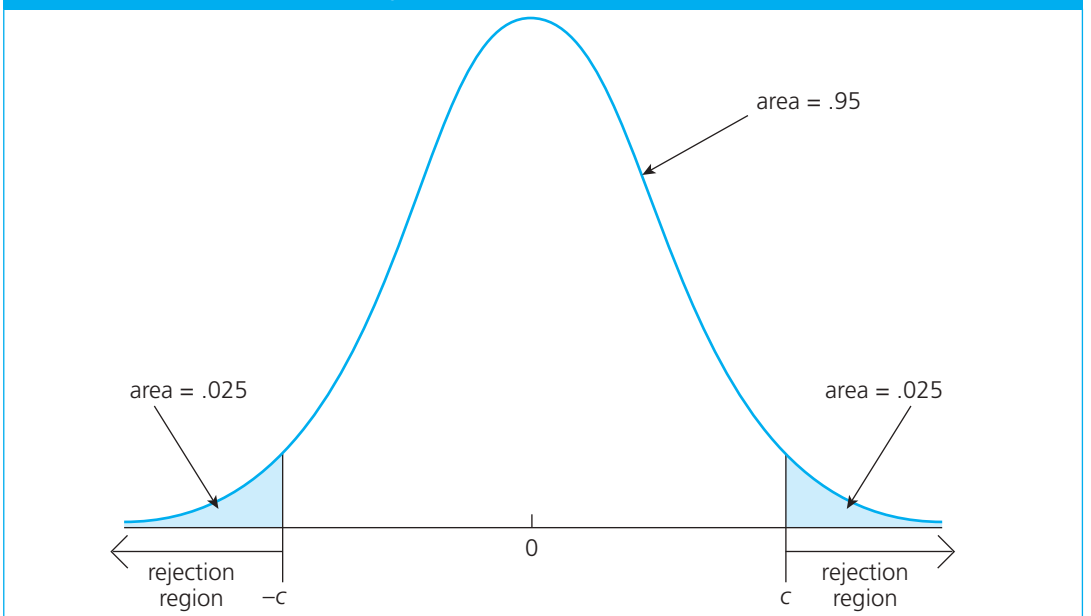
For two-sided alternatives, we must be careful to choose the critical value so that the significance level of the test is still α. If H_1 is given by H_1: $\mu \neq \mu_0$, then we reject H_0 if $\bar{y}$ is far from μ_0 in *absolute value*: a $\bar{y}$ much larger or much smaller than μ_0 provides evidence against H_0 in favor of H_1. A $100 \cdot \alpha\%$ level test is obtained from the rejection rule

$$|t| > c, \qquad [C.39]$$

where $|t|$ is the absolute value of the t statistic in (C.35). This gives a **two-tailed test**. We must now be careful in choosing the critical value: c is the $100(1 - \alpha/2)$ percentile in the t_{n-1} distribution. For example, if $\alpha = .05$, then the critical value is the 97.5^{th} percentile in the t_{n-1} distribution. This ensures that H_0 is rejected only 5% of the time when it is true (see Figure C.6). For example, if $n = 22$, then the critical value is $c = 2.08$, the 97.5^{th} percentile in a t_{21} distribution (see Table G.2). The absolute value of the t statistic must exceed 2.08 in order to reject H_0 against H_1 at the 5% level.

It is important to know the proper language of hypothesis testing. Sometimes, the appropriate phrase "we fail to reject H_0 in favor of H_1 at the 5% significance level" is replaced with "we accept H_0 at the 5% significance level." The latter wording is incorrect. With the same set of data, there are

FIGURE C.6 Rejection region for a 5% significance level test against the two-sided alternative H_1: $\mu \neq \mu_0$.



usually many hypotheses that cannot be rejected. In the earlier election example, it would be logically inconsistent to say that H_0: $\theta = .42$ and H_0: $\theta = .43$ are both "accepted," because only one of these can be true. But it is entirely possible that neither of these hypotheses is rejected. For this reason, we always say "fail to reject H_0" rather than "accept H_0."

C-6c Asymptotic Tests for Nonnormal Populations

If the sample size is large enough to invoke the central limit theorem (see Section C-3), the mechanics of hypothesis testing for population means are the *same* whether or not the population distribution is normal. The theoretical justification comes from the fact that, under the null hypothesis,

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S \overset{a}{\sim} \text{Normal}(0,1).$$

Therefore, with large n, we can compare the t statistic in (C.35) with the critical values from a standard normal distribution. Because the t_{n-1} distribution converges to the standard normal distribution as n gets large, the t and standard normal critical values will be very close for extremely large n. Because asymptotic theory is based on n increasing without bound, it cannot tell us whether the standard normal or t critical values are better. For moderate values of n, say, between 30 and 60, it is traditional to use the t distribution because we know this is correct for normal populations. For $n > 120$, the choice between the t and standard normal distributions is largely irrelevant because the critical values are practically the same.

Because the critical values chosen using either the standard normal or t distribution are only approximately valid for nonnormal populations, our chosen significance levels are also only approximate; thus, for nonnormal populations, our significance levels are really *asymptotic* significance levels. Thus, if we choose a 5% significance level, but our population is nonnormal, then the actual significance level will be larger or smaller than 5% (and we cannot know which is the case). When the sample size is large, the actual significance level will be very close to 5%. Practically speaking, the distinction is not important, so we will now drop the qualifier "asymptotic."

EXAMPLE C.5 Race Discrimination in Hiring

In the Urban Institute study of discrimination in hiring (see Example C.3) using the data in AUDIT, we are primarily interested in testing H_0: $\mu = 0$ against H_1: $\mu < 0$ where $\mu = \theta_B - \theta_W$ is the difference in probabilities that blacks and whites receive job offers. Recall that μ is the population mean of the variable $Y = B - W$, where B and W are binary indicators. Using the $n = 241$ paired comparisons in the data file AUDIT, we obtained $\bar{y} = -.133$ and $\text{se}(\bar{y}) = .482/\sqrt{241} \approx .031$. The t statistic for testing H_0: $\mu = 0$ is $t = -.133/.031 \approx -4.29$. You will remember from Math Refresher B that the standard normal distribution is, for practical purposes, indistinguishable from the t distribution with 240 degrees of freedom. The value -4.29 is so far out in the left tail of the distribution that we reject H_0 at any reasonable significance level. In fact, the .005 (one-half of a percent) critical value (for the one-sided test) is about -2.58. A t value of -4.29 is *very* strong evidence against H_0 in favor of H_1. Hence, we conclude that there is discrimination in hiring.

C-6d Computing and Using *p*-Values

The traditional requirement of choosing a significance level ahead of time means that different researchers, using the same data and same procedure to test the same hypothesis, could wind up with different conclusions. Reporting the significance level at which we are carrying out the test solves this problem to some degree, but it does not completely remove the problem.

To provide more information, we can ask the following question: What is the *largest* significance level at which we could carry out the test and still fail to reject the null hypothesis? This value is known as the ***p*-value** of a test (sometimes called the *prob-value*). Compared with choosing a significance level ahead of time and obtaining a critical value, computing a p-value is somewhat more difficult. But with the advent of quick and inexpensive computing, p-values are now fairly easy to obtain.

As an illustration, consider the problem of testing H_0: $\mu = 0$ in a Normal(μ, σ^2) population. Our test statistic in this case is $T = \sqrt{n}\cdot\overline{Y}/S$, and we assume that n is large enough to treat T as having a standard normal distribution under H_0. Suppose that the observed value of T for our sample is $t = 1.52$. (Note how we have skipped the step of choosing a significance level.) Now that we have seen the value t, we can find the largest significance level at which we would fail to reject H_0. This is the significance level associated with using t as our critical value. Because our test statistic T has a standard normal distribution under H_0, we have

$$p\text{-value} = P(T > 1.52|H_0) = 1 - \Phi(1.52) = .065, \qquad \text{[C.40]}$$

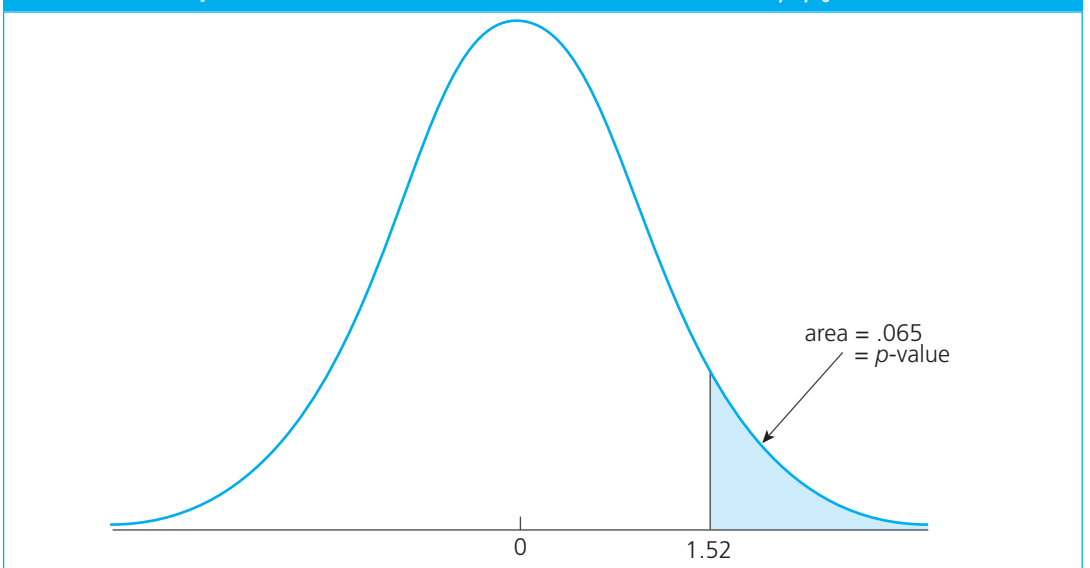
where $\Phi(\cdot)$ denotes the standard normal cdf. In other words, the p-value in this example is simply the area to the right of 1.52, the observed value of the test statistic, in a standard normal distribution. See Figure C.7 for illustration.

Because the p-value $= .065$, the largest significance level at which we can carry out this test and fail to reject is 6.5%. If we carry out the test at a level below 6.5% (such as at 5%), we fail to reject H_0. If we carry out the test at a level larger than 6.5% (such as 10%), we reject H_0. With the p-value at hand, we can carry out the test at any level.

The p-value in this example has another useful interpretation: it is the probability that we observe a value of T as large as 1.52 when the null hypothesis is true. If the null hypothesis is actually true, we would observe a value of T as large as 1.52 due to chance only 6.5% of the time. Whether this is small enough to reject H_0 depends on our tolerance for a Type I error. The p-value has a similar interpretation in all other cases, as we will see.

Generally, small p-values are evidence *against* H_0, because they indicate that the outcome of the data occurs with small probability if H_0 is true. In the previous example, if t had been a larger value, say, $t = 2.85$, then the p-value would be $1 - \Phi(2.85) \approx .002$. This means that, if the null hypothesis were true, we would observe a value of T as large as 2.85 with probability .002. How do we

FIGURE C.7 The *p*-value when $t = 1.52$ for the one-sided alternative $\mu\cdot\mu_0$.



interpret this? Either we obtained a very unusual sample or the null hypothesis is false. Unless we have a *very* small tolerance for Type I error, we would reject the null hypothesis. On the other hand, a large p-value is weak evidence against H_0. If we had gotten $t = .47$ in the previous example, then the p-value $= 1 - \Phi(.47) = .32$. Observing a value of T larger than .47 happens with probability .32, even when H_0 is true; this is large enough so that there is insufficient doubt about H_0, unless we have a very high tolerance for Type I error.

For hypothesis testing about a population mean using the t distribution, we need detailed tables in order to compute p-values. Table G.2 only allows us to put bounds on p-values. Fortunately, many statistics and econometrics packages now compute p-values routinely, and they also provide calculation of cdfs for the t and other distributions used for computing p-values.

EXAMPLE C.6 **Effect of Job Training Grants on Worker Productivity**

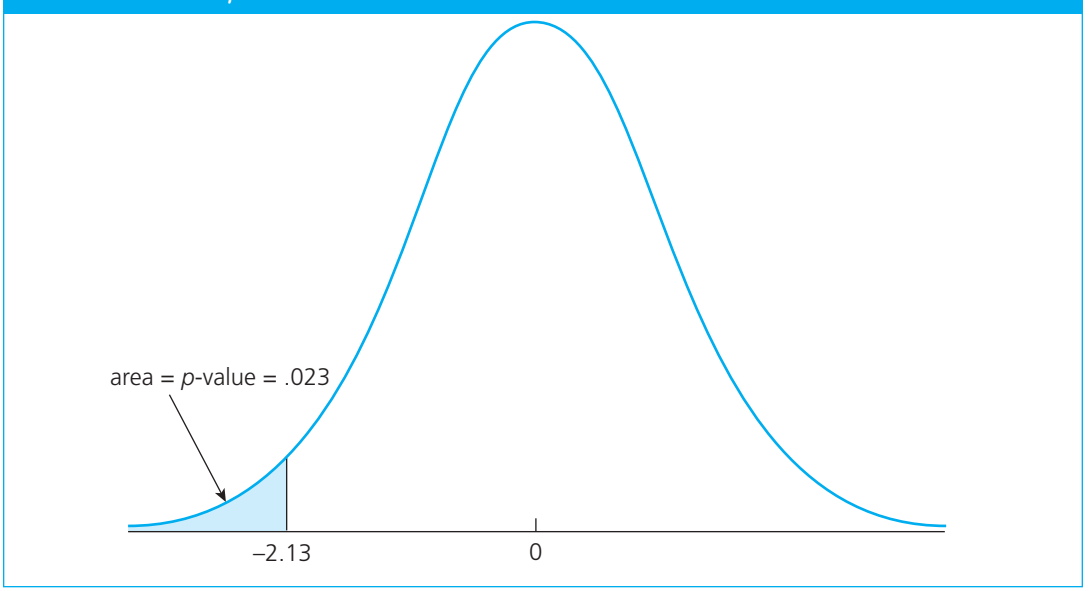
Consider again the Holzer et al. (1993) data in Example C.2. From a policy perspective, there are two questions of interest. First, what is our best estimate of the mean change in scrap rates, μ? We have already obtained this for the sample of 20 firms listed in Table C.3: the sample average of the change in scrap rates is -1.15. Relative to the initial average scrap rate in 1987, this represents a fall in the scrap rate of about 26.3% $(-1.15/4.38 \approx -.263)$, which is a nontrivial effect.

We would also like to know whether the sample provides strong evidence for an effect in the population of manufacturing firms that could have received grants. The null hypothesis is H_0: $\mu = 0$, and we test this against H_1: $\mu < 0$, where μ is the average change in scrap rates. Under the null, the job training grants have no effect on average scrap rates. The alternative states that there is an effect. We do not care about the alternative $\mu > 0$, so the null hypothesis is effectively H_0: $\mu \geq 0$.

Because $\bar{y} = -1.15$ and $\text{se}(\bar{y}) = .54$, $t = -1.15/.54 = -2.13$. This is below the 5% critical value of -1.73 (from a t_{19} distribution) but above the 1% critical value, -2.54. The p-value in this case is computed as

$$p\text{-value} = \text{P}(T_{19} < -2.13), \qquad \text{[C.41]}$$

FIGURE C.8 **The p-value when $t = -2.13$ with 19 degrees of freedom for the one-sided alternative $\mu < 0$.**



where T_{19} represents a t distributed random variable with 19 degrees of freedom. The inequality is reversed from (C.40) because the alternative has the form in (C.33). The probability in (C.41) is the area to the left of -2.13 in a t_{19} distribution (see Figure C.8).

Using Table G.2, the most we can say is that the p-value is between .025 and .01, but it is closer to .025 (because the 97.5th percentile is about 2.09). Using a statistical package, such as Stata®, we can compute the exact p-value. It turns out to be about .023, which is reasonable evidence against H_0. This is certainly enough evidence to reject the null hypothesis that the training grants had no effect at the 2.5% significance level (and therefore at the 5% level).

Computing a p-value for a two-sided test is similar, but we must account for the two-sided nature of the rejection rule. For t testing about population means, the p-value is computed as

$$P(|T_{n-1}| > |t|) = 2P(T_{n-1} > |t|), \qquad \text{[C.42]}$$

where t is the value of the test statistic and T_{n-1} is a t random variable. (For large n, replace T_{n-1} with a standard normal random variable.) Thus, compute the absolute value of the t statistic, find the area to the right of this value in a t_{n-1} distribution, and multiply the area by two.

For nonnormal populations, the exact p-value can be difficult to obtain. Nevertheless, we can find *asymptotic* p-values by using the same calculations. These p-values are valid for large sample sizes. For n larger than, say, 120, we might as well use the standard normal distribution. Table G.1 is detailed enough to get accurate p-values, but we can also use a statistics or econometrics program.

EXAMPLE C.7 **Race Discrimination in Hiring**

Using the matched pairs data from the Urban Institute in the AUDIT data file ($n = 241$), we obtained $t = -4.29$. If Z is a standard normal random variable, $P(Z < -4.29)$ is, for practical purposes, zero. In other words, the (asymptotic) p-value for this example is essentially zero. This is very strong evidence against H_0.

Summary of How to Use *p*-Values:

(i) Choose a test statistic T and decide on the nature of the alternative. This determines whether the rejection rule is $t > c$, $t < -c$, or $|t| > c$.

(ii) Use the observed value of the t statistic as the critical value and compute the corresponding significance level of the test. This is the p-value. If the rejection rule is of the form $t > c$, then $p\text{-value} = P(T > t)$. If the rejection rule is $t < -c$, then $p\text{-value} = P(T < t)$; if the rejection rule is $|t| > c$, then $p\text{-value} = P(|T| > |t|)$.

(iii) If a significance level α has been chosen, then we reject H_0 at the $100{\cdot}\alpha\%$ level if $p\text{-value} < \alpha$. If $p\text{-value} \geq \alpha$, then we fail to reject H_0 at the $100{\cdot}\alpha\%$ level. Therefore, it is a small p-value that leads to rejection of the null hypothesis.

C-6e The Relationship between Confidence Intervals and Hypothesis Testing

Because constructing confidence intervals and hypothesis tests both involve probability statements, it is natural to think that they are somehow linked. It turns out that they are. After a confidence interval has been constructed, we can carry out a variety of hypothesis tests.

The confidence intervals we have discussed are all two-sided by nature. (In this text, we will have no need to construct one-sided confidence intervals.) Thus, confidence intervals can be used to

test against *two-sided* alternatives. In the case of a population mean, the null is given by (C.31), and the alternative is (C.34). Suppose we have constructed a 95% confidence interval for μ. Then, if the hypothesized value of μ under H_0, μ_0, is not in the confidence interval, then H_0: $\mu = \mu_0$ is rejected against H_1: $\mu \neq \mu_0$ at the 5% level. If μ_0 lies in this interval, then we fail to reject H_0 at the 5% level. Notice how any value for μ_0 can be tested once a confidence interval is constructed, and because a confidence interval contains more than one value, there are many null hypotheses that will not be rejected.

EXAMPLE C.8 Training Grants and Worker Productivity

In the Holzer et al. example, we constructed a 95% confidence interval for the mean change in scrap rate μ as $[-2.28, -.02]$. Because zero is excluded from this interval, we reject H_0: $\mu = 0$ against H_1: $\mu \neq 0$ at the 5% level. This 95% confidence interval also means that we fail to reject H_0: $\mu = -2$ at the 5% level. In fact, there is a continuum of null hypotheses that are not rejected given this confidence interval.

C-6f Practical versus Statistical Significance

In the examples covered so far, we have produced three kinds of evidence concerning population parameters: point estimates, confidence intervals, and hypothesis tests. These tools for learning about population parameters are equally important. There is an understandable tendency for students to focus on confidence intervals and hypothesis tests because these are things to which we can attach confidence or significance levels. But in any study, we must also interpret the *magnitudes* of point estimates.

The sign and magnitude of $\bar{y}$ determine its **practical significance** and allow us to discuss the direction of an intervention or policy effect, and whether the estimated effect is "large" or "small." On the other hand, **statistical significance** of $\bar{y}$ depends on the magnitude of its t statistic. For testing H_0: $\mu = 0$, the t statistic is simply $t = \bar{y}/\text{se}(\bar{y})$. In other words, statistical significance depends on the ratio of $\bar{y}$ to its standard error. Consequently, a t statistic can be large because $\bar{y}$ is large or $\text{se}(\bar{y})$ is small. In applications, it is important to discuss both practical and statistical significance, being aware that an estimate can be statistically significant without being especially large in a practical sense. Whether an estimate is practically important depends on the context as well as on one's judgment, so there are no set rules for determining practical significance.

EXAMPLE C.9 Effect of Freeway Width on Commute Time

Let Y denote the change in commute time, measured in minutes, for commuters in a metropolitan area from before a freeway was widened to after the freeway was widened. Assume that $Y \sim \text{Normal}(\mu,\sigma^2)$. The null hypothesis that the widening did not reduce average commute time is H_0: $\mu = 0$; the alternative that it reduced average commute time is H_1: $\mu < 0$. Suppose a random sample of commuters of size $n = 900$ is obtained to determine the effectiveness of the freeway project. The average change in commute time is computed to be $\bar{y} = -3.6$, and the sample standard deviation is $s = 32.7$; thus, $\text{se}(\bar{y}) = 32.7/\sqrt{900} = 1.09$. The t statistic is $t = -3.6/1.09 \approx -3.30$, which is very statistically significant; the p-value is about .0005. Thus, we conclude that the freeway widening had a statistically significant effect on average commute time.

If the outcome of the hypothesis test is all that were reported from the study, it would be misleading. Reporting only statistical significance masks the fact that the estimated reduction in average commute time, 3.6 minutes, seems pretty meager, although this depends to some extent on what the average commute time was prior to widening the freeway. To be up front, we should report the point estimate of -3.6, along with the significance test.

Finding point estimates that are statistically significant without being practically significant can occur when we are working with large samples. To discuss why this happens, it is useful to have the following definition.

Test Consistency. A **consistent test** rejects H_0 with probability approaching one as the sample size grows whenever H_1 is true.

Another way to say that a test is consistent is that, as the sample size tends to infinity, the power of the test gets closer and closer to unity whenever H_1 is true. All of the tests we cover in this text have this property. In the case of testing hypotheses about a population mean, test consistency follows because the variance of $\overline{Y}$ converges to zero as the sample size gets large. The t statistic for testing H_0: $\mu = 0$ is $T = \overline{Y}/(S/\sqrt{n})$. Because $\text{plim}(\overline{Y}) = \mu$ and $\text{plim}(S) = \sigma$, it follows that if, say, $\mu > 0$, then T gets larger and larger (with high probability) as $n \to \infty$. In other words, no matter how close m is to zero, we can be almost certain to reject H_0: $\mu = 0$ given a large enough sample size. This says nothing about whether μ is large in a practical sense.

C-7 Remarks on Notation

In our review of probability and statistics here and in Math Refresher B, we have been careful to use standard conventions to denote random variables, estimators, and test statistics. For example, we have used W to indicate an estimator (random variable) and w to denote a particular estimate (outcome of the random variable W). Distinguishing between an estimator and an estimate is important for understanding various concepts in estimation and hypothesis testing. However, making this distinction quickly becomes a burden in econometric analysis because the models are more complicated: many random variables and parameters will be involved, and being true to the usual conventions from probability and statistics requires many extra symbols.

In the main text, we use a simpler convention that is widely used in econometrics. If θ is a population parameter, the notation $\hat{\theta}$ ("theta hat") will be used to denote both an estimator and an estimate of θ. This notation is useful in that it provides a simple way of attaching an estimator to the population parameter it is supposed to be estimating. Thus, if the population parameter is β, then $\hat{\beta}$ denotes an estimator or estimate of β; if the parameter is σ^2, $\hat{\sigma}^2$ is an estimator or estimate of σ^2; and so on. Sometimes, we will discuss two estimators of the same parameter, in which case we will need a different notation, such as $\tilde{\theta}$ ("theta tilde").

Although dropping the conventions from probability and statistics to indicate estimators, random variables, and test statistics puts additional responsibility on you, it is not a big deal once the difference between an estimator and an estimate is understood. If we are discussing *statistical* properties of $\hat{\theta}$—such as deriving whether or not it is unbiased or consistent—then we are necessarily viewing $\hat{\theta}$ as an estimator. On the other hand, if we write something like $\hat{\theta} = 1.73$, then we are clearly denoting a point estimate from a given sample of data. The confusion that can arise by using $\hat{\theta}$ to denote both should be minimal once you have a good understanding of probability and statistics.

Summary

We have discussed topics from mathematical statistics that are heavily relied upon in econometric analysis. The notion of an estimator, which is simply a rule for combining data to estimate a population parameter, is fundamental. We have covered various properties of estimators. The most important small sample properties are unbiasedness and efficiency, the latter of which depends on comparing variances when estimators are unbiased. Large sample properties concern the sequence of estimators

obtained as the sample size grows, and they are also depended upon in econometrics. Any useful estimator is consistent. The central limit theorem implies that, in large samples, the sampling distribution of most estimators is approximately normal.

The sampling distribution of an estimator can be used to construct confidence intervals. We saw this for estimating the mean from a normal distribution and for computing approximate confidence intervals in nonnormal cases. Classical hypothesis testing, which requires specifying a null hypothesis, an alternative hypothesis, and a significance level, is carried out by comparing a test statistic to a critical value. Alternatively, a *p*-value can be computed that allows us to carry out a test at any significance level.

Key Terms

Alternative Hypothesis
Asymptotic Normality
Bias
Biased Estimator
Central Limit Theorem (CLT)
Confidence Interval
Consistent Estimator
Consistent Test
Critical Value
Estimate
Estimator
Hypothesis Test
Inconsistent
Interval Estimator
Law of Large Numbers (LLN)
Least Squares Estimator
Log-Likelihood Function
Maximum Likelihood Estimator
Mean Squared Error (MSE)
Method of Moments
Minimum Variance Unbiased Estimator
Null Hypothesis
One-Sided Alternative
One-Tailed Test
Population
Power of a Test
Practical Significance
Probability Limit
p-Value
Random Sample
Rejection Region
Sample Average
Sample Correlation Coefficient
Sample Covariance
Sample Standard Deviation
Sample Variance
Sampling Distribution
Sampling Standard Deviation
Sampling Variance
Significance Level
Standard Error
Statistical Significance
t Statistic
Test Statistic
Two-Sided Alternative
Two-Tailed Test
Type I Error
Type II Error
Unbiased Estimator

Problems

1 Let Y_1, Y_2, Y_3, and Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\bar{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$ denote the average of these four random variables.

(i) What are the expected value and variance of $\bar{Y}$ in terms of μ and σ^2?

(ii) Now, consider a different estimator of μ:

$$W = \frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4.$$

This is an example of a *weighted* average of the Y_i. Show that W is also an unbiased estimator of μ. Find the variance of W.

(iii) Based on your answers to parts (i) and (ii), which estimator of m do you prefer, $\bar{Y}$ or W?

2 This is a more general version of Problem C.1. Let $Y_1, Y_2, \ldots, Y_n$ be n pairwise uncorrelated random variables with common mean m and common variance σ^2. Let $\bar{Y}$ denote the sample average.

(i) Define the class of *linear estimators* of μ by

$$W_a = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n,$$

where the a_i are constants. What restriction on the a_i is needed for W_a to be an unbiased estimator of μ?

(ii) Find $\text{Var}(W_a)$.

(iii) For any numbers $a_1, a_2, \ldots, a_n$, the following inequality holds: $(a_1 + a_2 + \cdots + a_n)^2/n \leq a_1^2 + a_2^2 + \cdots + a_n^2$. Use this, along with parts (i) and (ii), to show that $\text{Var}(W_a) \geq \text{Var}(\overline{Y})$ whenever W_a is unbiased, so that $\overline{Y}$ is the *best linear unbiased estimator*. [*Hint*: What does the inequality become when the a_i satisfy the restriction from part (i)?]

3 Let $\overline{Y}$ denote the sample average from a random sample with mean μ and variance σ^2. Consider two alternative estimators of μ: $W_1 = [(n-1)/n]\overline{Y}$ and $W_2 = \overline{Y}/2$.

(i) Show that W_1 and W_2 are both biased estimators of μ and find the biases. What happens to the biases as $n \to \infty$? Comment on any important differences in bias for the two estimators as the sample size gets large.

(ii) Find the probability limits of W_1 and W_2. {*Hint*: Use Properties PLIM.1 and PLIM.2; for W_1, note that plim $[(n-1)/n] = 1$.} Which estimator is consistent?

(iii) Find $\text{Var}(W_1)$ and $\text{Var}(W_2)$.

(iv) Argue that W_1 is a better estimator than $\overline{Y}$ if μ is "close" to zero. (Consider both bias and variance.)

4 For positive random variables X and Y, suppose the expected value of Y given X is $\text{E}(Y|X) = \theta X$. The unknown parameter θ shows how the expected value of Y changes with X.

(i) Define the random variable $Z = Y/X$. Show that $\text{E}(Z) = \theta$. [*Hint*: Use Property CE.2 in Math Refresher B along with the law of iterated expectations, Property CE.4 (also in Math Refresher B). In particular, first show that $\text{E}(Z|X) = \theta$ and then use CE.4.]

(ii) Use part (i) to prove that the estimator $W_1 = n^{-1}\sum_{i=1}^{n}(Y_i/X_i)$ is unbiased for θ, where $\{(X_i, Y_i): i = 1, 2, \ldots, n\}$ is a random sample.

(iii) Explain why the estimator $W_2 = \overline{Y}/\overline{X}$, where the overbars denote sample averages, is not the same as W_1. Nevertheless, show that W_2 is also unbiased for θ.

(iv) The following table contains data on corn yields for several counties in Iowa. The USDA predicts the number of hectares of corn in each county based on satellite photos. Researchers count the number of "pixels" of corn in the satellite picture (as opposed to, for example, the number of pixels of soybeans or of uncultivated land) and use these to predict the actual number of hectares. To develop a prediction equation to be used for counties in general, the USDA surveyed farmers in selected counties to obtain corn yields in hectares. Let Y_i = corn yield in county i and let X_i = number of corn pixels in the satellite picture for county i. There are $n = 17$ observations for eight counties. Use this sample to compute the estimates of θ devised in parts (ii) and (iii). Are the estimates similar?

Plot	Corn Yield	Corn Pixels
1	165.76	374
2	96.32	209
3	76.08	253
4	185.35	432
5	116.43	367
6	162.08	361
7	152.04	288
8	161.75	369
9	92.88	206
10	149.94	316
11	64.75	145
12	127.07	355
13	133.55	295
14	77.70	223
15	206.39	459
16	108.33	290
17	118.17	307

5 Let Y denote a Bernoulli(θ) random variable with $0 < \theta < 1$. Suppose we are interested in estimating the *odds ratio*, $\gamma = \theta/(1 - \theta)$, which is the probability of success over the probability of failure. Given a random sample $\{Y_1, \ldots, Y_n\}$, we know that an unbiased and consistent estimator of θ is $\overline{Y}$, the proportion of successes in n trials. A natural estimator of γ is $G = \overline{Y}/(1 - \overline{Y})$, the proportion of successes over the proportion of failures in the sample.

(i) Why is G not an unbiased estimator of γ?

(ii) Use PLIM.2 (iii) to show that G is a consistent estimator of γ.

6 You are hired by the governor to study whether a tax on liquor has decreased average liquor consumption in your state. You are able to obtain, for a sample of individuals selected at random, the difference in liquor consumption (in ounces) for the years before and after the tax. For person i who is sampled randomly from the population, Y_i denotes the change in liquor consumption. Treat these as a random sample from a Normal(μ, σ^2) distribution.

(i) The null hypothesis is that there was no change in average liquor consumption. State this formally in terms of μ.

(ii) The alternative is that there was a decline in liquor consumption; state the alternative in terms of μ.

(iii) Now, suppose your sample size is $n = 900$ and you obtain the estimates $\bar{y} = -32.8$ and $s = 466.4$. Calculate the t statistic for testing H_0 against H_1; obtain the p-value for the test. (Because of the large sample size, just use the standard normal distribution tabulated in Table G.1.) Do you reject H_0 at the 5% level? At the 1% level?

(iv) Would you say that the estimated fall in consumption is large in magnitude? Comment on the practical versus statistical significance of this estimate.

(v) What has been implicitly assumed in your analysis about other determinants of liquor consumption over the two-year period in order to infer causality from the tax change to liquor consumption?

7 The new management at a bakery claims that workers are now more productive than they were under old management, which is why wages have "generally increased." Let W_i^b be Worker i's wage under the old management and let W_i^a be Worker i's wage after the change. The difference is $D_i \equiv W_i^a - W_i^b$. Assume that the D_i are a random sample from a Normal (μ, σ^2) distribution.

(i) Using the following data on 15 workers, construct an exact 95% confidence interval for μ.

(ii) Formally state the null hypothesis that there has been no change in average wages. In particular, what is $E(D_i)$ under H_0? If you are hired to examine the validity of the new management's claim, what is the relevant alternative hypothesis in terms of $\mu = E(D_i)$?

(iii) Test the null hypothesis from part (ii) against the stated alternative at the 5% and 1% levels.

(iv) Obtain the p-value for the test in part (iii).

Worker	Wage Before	Wage After
1	8.30	9.25
2	9.40	9.00
3	9.00	9.25
4	10.50	10.00
5	11.40	12.00
6	8.75	9.50
7	10.00	10.25
8	9.50	9.50
9	10.80	11.50
10	12.55	13.10
11	12.00	11.50
12	8.65	9.00
13	7.75	7.75
14	11.25	11.50
15	12.65	13.00

8 The *New York Times* (2/5/90) reported three-point shooting performance for the top 10 three-point shooters in the NBA. The following table summarizes these data:

Player	FGA-FGM
Mark Price	429-188
Trent Tucker	833-345
Dale Ellis	1,149-472
Craig Hodges	1,016-396
Danny Ainge	1,051-406
Byron Scott	676-260
Reggie Miller	416-159
Larry Bird	1,206-455
Jon Sundvold	440-166
Brian Taylor	417-157

Note: FGA = field goals attempted and FGM = field goals made.

For a given player, the outcome of a particular shot can be modeled as a Bernoulli (zero-one) variable: if Y_i is the outcome of shot i, then $Y_i = 1$ if the shot is made, and $Y_i = 0$ if the shot is missed. Let θ denote the probability of making any particular three-point shot attempt. The natural estimator of θ is $\overline{Y} = FGM/FGA$.

(i) Estimate θ for Mark Price.

(ii) Find the standard deviation of the estimator $\overline{Y}$ in terms of θ and the number of shot attempts, n.

(iii) The asymptotic distribution of $(\overline{Y} - \theta)/\text{se}(\overline{Y})$ is standard normal, where $\text{se}(\overline{Y}) = \sqrt{\overline{Y}(1 - \overline{Y})/n}$. Use this fact to test H_0: $\theta = .5$ against H_1: $\theta < .5$ for Mark Price. Use a 1% significance level.

9 Suppose that a military dictator in an unnamed country holds a plebiscite (a yes/no vote of confidence) and claims that he was supported by 65% of the voters. A human rights group suspects foul play and hires you to test the validity of the dictator's claim. You have a budget that allows you to randomly sample 200 voters from the country.

(i) Let X be the number of yes votes obtained from a random sample of 200 out of the entire voting population. What is the expected value of X if, in fact, 65% of all voters supported the dictator?

(ii) What is the standard deviation of X, again assuming that the true fraction voting yes in the plebiscite is .65?

(iii) Now, you collect your sample of 200, and you find that 115 people actually voted yes. Use the CLT to approximate the probability that you would find 115 or fewer yes votes from a random sample of 200 if, in fact, 65% of the entire population voted yes.

(iv) How would you explain the relevance of the number in part (iii) to someone who does not have training in statistics?

10 Before a strike prematurely ended the 1994 major league baseball season, Tony Gwynn of the San Diego Padres had 165 hits in 419 at bats, for a .394 batting average. There was discussion about whether Gwynn was a potential .400 hitter that year. This issue can be couched in terms of Gwynn's probability of getting a hit on a particular at bat, call it θ. Let Y_i be the Bernoulli(θ) indicator equal to unity if Gwynn gets a hit during his i^{th} at bat, and zero otherwise. Then, $Y_1, Y_2, \ldots, Y_n$ is a random sample from a Bernoulli(θ) distribution, where θ is the probability of success, and $n = 419$.

Our best point estimate of θ is Gwynn's batting average, which is just the proportion of successes: $\bar{y} = .394$. Using the fact that $\text{se}(\bar{y}) = \sqrt{\bar{y}(1 - \bar{y})/n}$, construct an approximate 95% confidence interval for θ, using the standard normal distribution. Would you say there is strong evidence against Gwynn's being a potential .400 hitter? Explain.

11 Suppose that between their first and second years in college, 400 students are randomly selected and given a university grant to purchase a new computer. For student i, y_i denotes the change in GPA from the first year to the second year. If the average change is $\bar{y} = .132$ with standard deviation $s = 1.27$, is the average change in GPAs statistically greater than zero?

12 (Requires Calculus) A count random variable, say Y, takes on nonnegative integer values, $\{0, 1, 2, \ldots\}$. The most common distribution for a count variable is the *Poisson*(θ) distribution, where the parameter θ is the expected value: $\theta = \text{E}(Y)$. The probability density function is

$$f(y; \theta) = \exp(-\theta)\theta^y/y!, \; y = 0, 1, 2, \ldots$$
$$= 0 \text{ otherwise}$$

It can be shown that $\text{Var}(Y) = \theta$, so that the mean and variance are the same.

(i) For a random draw Y_i from the population, find the log-likelihood function $\ell(\theta; Y_i) = \log[f(Y_i; \theta)]$. What is the log likelihood for a random sample of size n, say $\mathcal{L}_n(\theta)$? [Hint: Look at equation (C.16).]

(ii) Using the notational convention in Section C.7, find the first order condition for the MLE, $\hat{\theta}$, and show that $\hat{\theta} = \bar{Y}$, the sample average.

(iii) Why is $\hat{\theta}$ unbiased?

(iv) Find $\text{Var}(\bar{Y})$ as a function of θ and n.

(v) Why is $\bar{Y}$ consistent?

(vi) Do the unbiasedness and consistency of the MLE in this case depend on whether the Poisson distribution is correct? Explain.

(vii) What is the distribution of

$$\frac{\sqrt{n}(\bar{Y} - \theta)}{\sqrt{\theta}}$$

as $n \to \infty$? Explain.

(viii) If $\text{E}(Y) = \theta$ but $\text{Var}(Y) = v(\theta) > 0$—so that the Poisson distribution may fail—modify the random variable in (vii) so that it has a limiting distribution that does not depend on θ.

Advanced Treatment D

Summary of Matrix Algebra

This Advanced Treatment summarizes the matrix algebra concepts, including the algebra of probability, needed for the study of multiple linear regression models using matrices in Advanced Treatment E. None of this material is used in the main text.

D-1 Basic Definitions

Definition D.1 (Matrix). A **matrix** is a rectangular array of numbers. More precisely, an $m \times n$ matrix has m rows and n columns. The positive integer m is called the *row dimension*, and n is called the *column dimension*.

We use uppercase boldface letters to denote matrices. We can write an $m \times n$ matrix generically as

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & & & & \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix},$$

where a_{ij} represents the element in the i^{th} row and the j^{th} column. For example, a_{25} stands for the number in the second row and the fifth column of $\mathbf{A}$. A specific example of a 2×3 matrix is

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \qquad \text{[D.1]}$$

where $a_{13} = 7$. The shorthand $\mathbf{A} = [a_{ij}]$ is often used to define matrix operations.

Definition D.2 (Square Matrix). A **square matrix** has the same number of rows and columns. The dimension of a square matrix is its number of rows and columns.

Definition D.3 (Vectors)

(i) A $1 \times m$ matrix is called a **row vector** (of dimension m) and can be written as $\mathbf{x} \equiv (x_1, x_2, \dots, x_m)$.

(ii) An $n \times 1$ matrix is called a **column vector** and can be written as

$$\mathbf{y} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Definition D.4 (Diagonal Matrix). A square matrix $\mathbf{A}$ is a **diagonal matrix** when all of its off-diagonal elements are zero, that is, $a_{ij} = 0$ for all $i \neq j$. We can always write a diagonal matrix as

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix}.$$

Definition D.5 (Identity and Zero Matrices)

(i) The $n \times n$ **identity matrix**, denoted $\mathbf{I}$, or sometimes $\mathbf{I}_n$ to emphasize its dimension, is the diagonal matrix with unity (one) in each diagonal position, and zero elsewhere:

$$\mathbf{I} \equiv \mathbf{I}_n \equiv \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

(ii) The $m \times n$ **zero matrix**, denoted $\mathbf{0}$, is the $m \times n$ matrix with zero for all entries. This need not be a square matrix.

D-2 Matrix Operations

D-2a Matrix Addition

Two matrices $\mathbf{A}$ and $\mathbf{B}$, each having dimension $m \times n$, can be added element by element: $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$. More precisely,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{bmatrix}.$$

For example,

$$\begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -4 \\ 4 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 3 & -1 & 3 \\ 0 & 7 & 3 \end{bmatrix}.$$

Matrices of different dimensions cannot be added.

D-2b Scalar Multiplication

Given any real number γ (often called a scalar), **scalar multiplication** is defined as $\gamma\mathbf{A} \equiv [\gamma a_{ij}]$, or

$$\gamma\mathbf{A} = \begin{bmatrix} \gamma a_{11} & \gamma a_{12} & \dots & \gamma a_{1n} \\ \gamma a_{21} & \gamma a_{22} & \dots & \gamma a_{2n} \\ \vdots & & & \\ \gamma a_{m1} & \gamma a_{m2} & \dots & \gamma a_{mn} \end{bmatrix}.$$

For example, if $\gamma = 2$ and $\mathbf{A}$ is the matrix in equation (D.1), then

$$\gamma\mathbf{A} = \begin{bmatrix} 4 & -2 & 14 \\ -8 & 10 & 0 \end{bmatrix}.$$

D-2c Matrix Multiplication

To multiply matrix $\mathbf{A}$ by matrix $\mathbf{B}$ to form the product $\mathbf{AB}$, the *column* dimension of $\mathbf{A}$ must equal the *row* dimension of $\mathbf{B}$. Therefore, let $\mathbf{A}$ be an $m \times n$ matrix and let $\mathbf{B}$ be an $n \times p$ matrix. Then, **matrix multiplication** is defined as

$$\mathbf{AB} = \left[\sum_{k=1}^{n} a_{ik} b_{kj} \right].$$

In other words, the $(i, j)^{\text{th}}$ element of the new matrix $\mathbf{AB}$ is obtained by multiplying each element in the i^{th} row of $\mathbf{A}$ by the corresponding element in the j^{th} column of $\mathbf{B}$ and adding these n products together. A schematic may help make this process more transparent:

$$\begin{array}{ccccc} & \mathbf{A} & & \mathbf{B} & \mathbf{AB} \\ i^{\text{th}}\text{ row} \rightarrow & \begin{bmatrix} a_{i1} a_{i2} a_{i3} \ldots a_{in} \end{bmatrix} & \begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \\ \vdots \\ b_{nj} \end{bmatrix} & = & \begin{bmatrix} \sum_{k=1}^{n} a_{ik} b_{kj} \end{bmatrix} \\ & & \uparrow & & \uparrow \\ & & j^{\text{th}}\text{ column} & & (i, j)^{\text{th}}\text{ element} \end{array},$$

where, by the definition of the summation operator in Math Refresher A,

$$\sum_{k=1}^{n} a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj}.$$

For example,

$$\begin{bmatrix} 2 & -1 & 0 \\ -4 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 6 & 0 \\ -1 & 2 & 0 & 1 \\ 3 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 12 & -1 \\ -1 & -2 & -24 & 1 \end{bmatrix}.$$

We can also multiply a matrix and a vector. If $\mathbf{A}$ is an $n \times m$ matrix and $\mathbf{y}$ is an $m \times 1$ vector, then $\mathbf{Ay}$ is an $n \times 1$ vector. If $\mathbf{x}$ is a $1 \times n$ vector, then $\mathbf{xA}$ is a $1 \times m$ vector.

Matrix addition, scalar multiplication, and matrix multiplication can be combined in various ways, and these operations satisfy several rules that are familiar from basic operations on numbers. In the following list of properties, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are matrices with appropriate dimensions for applying each operation, and α and β are real numbers. Most of these properties are easy to illustrate from the definitions.

Properties of Matrix Operations. (1) $(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$; (2) $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$; (3) $(\alpha\beta)\mathbf{A} = \alpha(\beta\mathbf{A})$; (4) $\alpha(\mathbf{AB}) = (\alpha\mathbf{A})\mathbf{B}$; (5) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$; (6) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$; (7) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$; (8) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$; (9) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$; (10) $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$; (11) $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$; (12) $\mathbf{A} - \mathbf{A} = \mathbf{0}$; (13) $\mathbf{A0} = \mathbf{0A} = \mathbf{0}$; and (14) $\mathbf{AB} \neq \mathbf{BA}$, even when both products are defined.

The last property deserves further comment. If $\mathbf{A}$ is $n \times m$ and $\mathbf{B}$ is $m \times p$, then $\mathbf{AB}$ is defined, but $\mathbf{BA}$ is defined only if $n = p$ (the row dimension of $\mathbf{A}$ equals the column dimension of $\mathbf{B}$). If $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$, then $\mathbf{AB}$ and $\mathbf{BA}$ are both defined, but they are not usually the same; in fact, they have different dimensions, unless $\mathbf{A}$ and $\mathbf{B}$ are both square matrices. Even when $\mathbf{A}$ and $\mathbf{B}$ are both square, $\mathbf{AB} \neq \mathbf{BA}$, except under special circumstances.

D-2d Transpose

Definition D.6 (Transpose). Let $\mathbf{A} = [a_{ij}]$ be an $m \times n$ matrix. The **transpose** of $\mathbf{A}$, denoted $\mathbf{A}'$ (called $\mathbf{A}$ *prime*), is the $n \times m$ matrix obtained by interchanging the rows and columns of $\mathbf{A}$. We can write this as $\mathbf{A}' \equiv [a_{ji}]$.

For example,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \quad \mathbf{A}' = \begin{bmatrix} 2 & -4 \\ -1 & 5 \\ 7 & 0 \end{bmatrix}.$$

Properties of Transpose. (1) $(\mathbf{A}')' = \mathbf{A}$; (2) $(\alpha\mathbf{A})' = \alpha\mathbf{A}'$ for any scalar α; (3) $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$; (4) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, where $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times k$; (5) $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2$, where $\mathbf{x}$ is an $n \times 1$ vector; and (6) If $\mathbf{A}$ is an $n \times k$ matrix with *rows* given by the $1 \times k$ vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, so that we can write

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix},$$

then $\mathbf{A}' = (\mathbf{a}_1' \mathbf{a}_2' \ldots \mathbf{a}_n')$.

Definition D.7 (Symmetric Matrix). A square matrix $\mathbf{A}$ is a **symmetric matrix** if, and only if, $\mathbf{A}' = \mathbf{A}$.

If $\mathbf{X}$ is any $n \times k$ matrix, then $\mathbf{X}'\mathbf{X}$ is always defined and is a symmetric matrix, as can be seen by applying the first and fourth transpose properties (see Problem 3).

D-2e Partitioned Matrix Multiplication

Let $\mathbf{A}$ be an $n \times k$ matrix with rows given by the $1 \times k$ vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, and let $\mathbf{B}$ be an $n \times m$ matrix with rows given by $1 \times m$ vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}.$$

Then,

$$\mathbf{A}'\mathbf{B} = \sum_{i=1}^{n} \mathbf{a}_i' \mathbf{b}_i,$$

where for each i, $\mathbf{a}_i'\mathbf{b}_i$ is a $k \times m$ matrix. Therefore, $\mathbf{A}'\mathbf{B}$ can be written as the sum of n matrices, each of which is $k \times m$. As a special case, we have

$$\mathbf{A}'\mathbf{A} = \sum_{i=1}^{n} \mathbf{a}_i'\mathbf{a}_i,$$

where $\mathbf{a}_i'\mathbf{a}_i$ is a $k \times k$ matrix for all i.

A more general form of partitioned matrix multiplication holds when we have matrices $\mathbf{A}$ $(m \times n)$ and $\mathbf{B}$ $(n \times p)$ written as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11}$ is $m_1 \times n_1$, $\mathbf{A}_{12}$ is $m_1 \times n_2$, $\mathbf{A}_{21}$ is $m_2 \times n_1$, $\mathbf{A}_{22}$ is $m_2 \times n_2$, $\mathbf{B}_{11}$ is $n_1 \times p_1$, $\mathbf{B}_{12}$ is $n_1 \times p_2$, $\mathbf{B}_{21}$ is $n_2 \times p_1$, and $\mathbf{B}_{22}$ is $n_2 \times p_2$. Naturally, $m_1 + m_2 = m$, $n_1 + n_2 = n$, and $p_1 + p_2 = p$.

When we form the product **AB,** the expression looks just like when the entries are scalars:

$$\mathbf{AB} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix}.$$

Note that each of the matrix multiplications that form the partition on the right is well defined because the column and row dimensions are compatible for multiplication.

D-2f Trace

The trace of a matrix is a very simple operation defined only for *square* matrices.

Definition D.8 (Trace). For any $n \times n$ matrix $\mathbf{A}$, the **trace of a matrix A**, denoted $\text{tr}(\mathbf{A})$, is the sum of its diagonal elements. Mathematically,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

Properties of Trace. (1) $\text{tr}(\mathbf{I}_n) = n$; (2) $\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A})$; (3) $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$; (4) $\text{tr}(\alpha\mathbf{A}) = \alpha\text{tr}(\mathbf{A})$, for any scalar α; and (5) $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, where $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$.

D-2g Inverse

The notion of a matrix inverse is very important for square matrices.

Definition D.9 (Inverse). An $n \times n$ matrix $\mathbf{A}$ has an **inverse**, denoted $\mathbf{A}^{-1}$, provided that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$ and $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$. In this case, $\mathbf{A}$ is said to be *invertible* or *nonsingular.* Otherwise, it is said to be *noninvertible* or *singular*.

Properties of Inverse. (1) If an inverse exists, it is unique; (2) $(\alpha\mathbf{A})^{-1} = (1/\alpha)\mathbf{A}^{-1}$, if $\alpha \neq 0$ and $\mathbf{A}$ is invertible; (3) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, if $\mathbf{A}$ and $\mathbf{B}$ are both $n \times n$ and invertible; and (4) $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

We will not be concerned with the mechanics of calculating the inverse of a matrix. Any matrix algebra text contains detailed examples of such calculations.

D-3 Linear Independence and Rank of a Matrix

For a set of vectors having the same dimension, it is important to know whether one vector can be expressed as a linear combination of the remaining vectors.

Definition D.10 (Linear Independence). Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ be a set of $n \times 1$ vectors. These are **linearly independent vectors** if, and only if,

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_r\mathbf{x}_r = \mathbf{0} \qquad [D.2]$$

implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$. If (D.2) holds for a set of scalars that are not all zero, then $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ is *linearly dependent*.

The statement that $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ is linearly dependent is equivalent to saying that at least one vector in this set can be written as a linear combination of the others.

Definition D.11 (Rank)

(i) Let $\mathbf{A}$ be an $n \times m$ matrix. The **rank of a matrix A**, denoted rank($\mathbf{A}$), is the maximum number of linearly independent columns of $\mathbf{A}$.

(ii) If $\mathbf{A}$ is $n \times m$ and $\text{rank}(\mathbf{A}) = m$, then $\mathbf{A}$ has *full column rank*.

If $\mathbf{A}$ is $n \times m$, its rank can be at most m. A matrix has full column rank if its columns form a linearly independent set. For example, the 3×2 matrix

$$\begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 0 & 0 \end{bmatrix}$$

can have at most rank two. In fact, its rank is only one because the second column is three times the first column.

Properties of Rank. (1) $\text{rank}(\mathbf{A}') = \text{rank}(\mathbf{A})$; (2) If $\mathbf{A}$ is $n \times k$, then $\text{rank}(\mathbf{A}) \leq \min(n, k)$; and (3) If $\mathbf{A}$ is $k \times k$ and $\text{rank}(\mathbf{A}) = k$, then $\mathbf{A}$ is invertible.

D-4 Quadratic Forms and Positive Definite Matrices

Definition D.12 (Quadratic Form). Let $\mathbf{A}$ be an $n \times n$ *symmetric* matrix. The **quadratic form** associated with the matrix $\mathbf{A}$ is the real-valued function defined for all $n \times 1$ vectors $\mathbf{x}$:

$$f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j>i}^{n} a_{ij}x_ix_j.$$

Definition D.13 (Positive Definite and Positive Semi-Definite)

(i) A symmetric matrix $\mathbf{A}$ is said to be **positive definite (p.d.)** if

$$\mathbf{x}'\mathbf{A}\mathbf{x} > 0 \text{ for all } n \times 1 \text{ vectors } \mathbf{x} \text{ except } \mathbf{x} = \mathbf{0}.$$

(ii) A symmetric matrix $\mathbf{A}$ is **positive semi-definite (p.s.d.)** if

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0 \text{ for all } n \times 1 \text{ vectors.}$$

If a matrix is positive definite or positive semi-definite, it is automatically assumed to be symmetric.

Properties of Positive Definite and Positive Semi-Definite Matrices. (1) A p.d. matrix has diagonal elements that are strictly positive, while a p.s.d. matrix has nonnegative diagonal elements; (2) If $\mathbf{A}$ is p.d., then $\mathbf{A}^{-1}$ exists and is p.d.; (3) If $\mathbf{X}$ is $n \times k$, then $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ are p.s.d.; and (4) If $\mathbf{X}$ is $n \times k$ and $\text{rank}(\mathbf{X}) = k$, then $\mathbf{X}'\mathbf{X}$ is p.d. (and therefore nonsingular).

D-5 Idempotent Matrices

Definition D.14 (Idempotent Matrix). Let $\mathbf{A}$ be an $n \times n$ symmetric matrix. Then $\mathbf{A}$ is said to be an **idempotent matrix** if, and only if, $\mathbf{AA} = \mathbf{A}$.

For example,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

is an idempotent matrix, as direct multiplication verifies.

Properties of Idempotent Matrices. Let $\mathbf{A}$ be an $n \times n$ idempotent matrix. (1) $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A})$, and (2) $\mathbf{A}$ is positive semi-definite.

We can construct idempotent matrices very generally. Let $\mathbf{X}$ be an $n \times k$ matrix with $\text{rank}(\mathbf{X}) = k$. Define

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I}_n - \mathbf{P}.$$

Then $\mathbf{P}$ and $\mathbf{M}$ are symmetric, idempotent matrices with $\text{rank}(\mathbf{P}) = k$ and $\text{rank}(\mathbf{M}) = n - k$. The ranks are most easily obtained by using Property 1: $\text{tr}(\mathbf{P}) = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]$ (from Property 5 for trace) $= \text{tr}(\mathbf{I}_k) = k$ (by Property 1 for trace). It easily follows that $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{P}) = n - k$.

D-6 Differentiation of Linear and Quadratic Forms

For a given $n \times 1$ vector $\mathbf{a}$, consider the linear function defined by

$$f(\mathbf{x}) = \mathbf{a}'\mathbf{x},$$

for all $n \times 1$ vectors $\mathbf{x}$. The derivative of f with respect to $\mathbf{x}$ is the $1 \times n$ vector of partial derivatives, which is simply

$$\partial f(\mathbf{x})/\partial \mathbf{x} = \mathbf{a}'.$$

For an $n \times n$ symmetric matrix $\mathbf{A}$, define the quadratic form

$$g(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}.$$

Then,

$$\partial g(\mathbf{x})/\partial \mathbf{x} = 2\mathbf{x}'\mathbf{A},$$

which is a $1 \times n$ vector.

D-7 Moments and Distributions of Random Vectors

In order to derive the expected value and variance of the OLS estimators using matrices, we need to define the expected value and variance of a **random vector**. As its name suggests, a random vector is simply a vector of random variables. We also need to define the multivariate normal distribution. These concepts are simply extensions of those covered in Math Refresher B.

D-7a Expected Value

Definition D.15 (Expected Value)

(i) If $\mathbf{y}$ is an $n \times 1$ random vector, the **expected value** of $\mathbf{y}$, denoted $\mathrm{E}(\mathbf{y})$, is the vector of expected values: $\mathrm{E}(\mathbf{y}) = [\mathrm{E}(y_1), \mathrm{E}(y_2), \ldots, \mathrm{E}(y_n)]'$.

(ii) If $\mathbf{Z}$ is an $n \times m$ random matrix, $\mathrm{E}(\mathbf{Z})$ is the $n \times m$ matrix of expected values: $\mathrm{E}(\mathbf{Z}) = [\mathrm{E}(z_{ij})]$.

Properties of Expected Value. (1) If $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b}$ is an $n \times 1$ vector, where both are nonrandom, then $\mathrm{E}(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}\mathrm{E}(\mathbf{y}) + \mathbf{b}$; and (2) If $\mathbf{A}$ is $p \times n$ and $\mathbf{B}$ is $m \times k$, where both are nonrandom, then $\mathrm{E}(\mathbf{AZB}) = \mathbf{A}\mathrm{E}(\mathbf{Z})\mathbf{B}$.

D-7b Variance-Covariance Matrix

Definition D.16 (Variance-Covariance Matrix). If $\mathbf{y}$ is an $n \times 1$ random vector, its **variance-covariance matrix**, denoted $\mathrm{Var}(\mathbf{y})$, is defined as

$$\mathrm{Var}(\mathbf{y}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & & & \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix},$$

where $\sigma_j^2 = \mathrm{Var}(y_j)$ and $\sigma_{ij} = \mathrm{Cov}(y_i, y_j)$. In other words, the variance-covariance matrix has the variances of each element of $\mathbf{y}$ down its diagonal, with covariance terms in the off diagonals. Because $\mathrm{Cov}(y_i, y_j) = \mathrm{Cov}(y_j, y_i)$, it immediately follows that a variance-covariance matrix is symmetric.

Properties of Variance. (1) If $\mathbf{a}$ is an $n \times 1$ nonrandom vector, then $\mathrm{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'[\mathrm{Var}(\mathbf{y})]\mathbf{a} \geq 0$; (2) If $\mathrm{Var}(\mathbf{a}'\mathbf{y}) > 0$ for all $\mathbf{a} \neq \mathbf{0}$, $\mathrm{Var}(\mathbf{y})$ is positive definite; (3) $\mathrm{Var}(\mathbf{y}) = \mathrm{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})']$, where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$; (4) If the elements of $\mathbf{y}$ are uncorrelated, $\mathrm{Var}(\mathbf{y})$ is a diagonal matrix. If, in addition, $\mathrm{Var}(y_j) = \sigma^2$ for $j = 1, 2, \ldots, n$, then $\mathrm{Var}(\mathbf{y}) = \sigma^2\mathbf{I}_n$; and (5) If $\mathbf{A}$ is an $m \times n$ nonrandom matrix and $\mathbf{b}$ is an $n \times 1$ nonrandom vector, then $\mathrm{Var}(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}[\mathrm{Var}(\mathbf{y})]\mathbf{A}'$.

D-7c Multivariate Normal Distribution

The normal distribution for a random variable was discussed at some length in Math Refresher B. We need to extend the normal distribution to random vectors. We will not provide an expression for the probability distribution function, as we do not need it. It is important to know that a multivariate normal random vector is completely characterized by its mean and its variance-covariance matrix. Therefore, if $\mathbf{y}$ is an $n \times 1$ multivariate normal random vector with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, we write $\mathbf{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We now state several useful properties of the **multivariate normal distribution**.

Properties of the Multivariate Normal Distribution. (1) If $\mathbf{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then each element of $\mathbf{y}$ is normally distributed; (2) If $\mathbf{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then y_i and y_j, any two elements of $\mathbf{y}$, are independent if, and only if, they are uncorrelated, that is, $\sigma_{ij} = 0$; (3) If $\mathbf{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{Ay} + \mathbf{b} \sim \mathrm{Normal}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$, where $\mathbf{A}$ and $\mathbf{b}$ are nonrandom; (4) If $\mathbf{y} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{\Sigma})$,

then, for nonrandom matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{Ay}$ and $\mathbf{By}$ are independent if, and only if, $\mathbf{A\Sigma B}' = \mathbf{0}$. In particular, if $\mathbf{\Sigma} = \sigma^2\mathbf{I}_n$, then $\mathbf{AB}' = \mathbf{0}$ is necessary and sufficient for independence of $\mathbf{Ay}$ and $\mathbf{By}$; (5) If $\mathbf{y} \sim \text{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$, $\mathbf{A}$ is a $k \times n$ nonrandom matrix, and $\mathbf{B}$ is an $n \times n$ symmetric, idempotent matrix, then $\mathbf{Ay}$ and $\mathbf{y}'\mathbf{By}$ are independent if, and only if, $\mathbf{AB} = \mathbf{0}$; and (6) If $\mathbf{y} \sim \text{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are nonrandom symmetric, idempotent matrices, then $\mathbf{y}'\mathbf{Ay}$ and $\mathbf{y}'\mathbf{By}$ are independent if, and only if, $\mathbf{AB} = \mathbf{0}$.

D-7d Chi-Square Distribution

In Math Refresher B, we defined a **chi-square random variable** as the sum of *squared* independent standard normal random variables. In vector notation, if $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$, then $\mathbf{u}'\mathbf{u} \sim \chi^2_n$.

Properties of the Chi-Square Distribution. (1) If $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ is an $n \times n$ symmetric, idempotent matrix with $\text{rank}(\mathbf{A}) = q$, then $\mathbf{u}'\mathbf{Au} \sim \chi^2_q$; (2) If $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ symmetric, idempotent matrices such that $\mathbf{AB} = \mathbf{0}$, then $\mathbf{u}'\mathbf{Au}$ and $\mathbf{u}'\mathbf{Bu}$ are independent, chi-square random variables; and (3) If $\mathbf{z} \sim \text{Normal}(\mathbf{0},\mathbf{C})$, where $\mathbf{C}$ is an $m \times m$ nonsingular matrix, then $\mathbf{z}'\mathbf{C}^{-1}\mathbf{z} \sim \chi^2_m$.

D-7e *t* Distribution

We also defined the ***t* distribution** in Math Refresher B. Now we add an important property.

Property of the t Distribution. If $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$, $\mathbf{c}$ is an $n \times 1$ nonrandom vector, $\mathbf{A}$ is a nonrandom $n \times n$ symmetric, idempotent matrix with rank q, and $\mathbf{Ac} = \mathbf{0}$, then $\{\mathbf{c}'\mathbf{u}/(\mathbf{c}'\mathbf{c})^{1/2}\}/(\mathbf{u}'\mathbf{Au}/q)^{1/2} \sim t_q$.

D-7f *F* Distribution

Recall that an ***F* random variable** is obtained by taking two *independent* chi-square random variables and finding the ratio of each, standardized by degrees of freedom.

Property of the F Distribution. If $\mathbf{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ nonrandom symmetric, idempotent matrices with $\text{rank}(\mathbf{A}) = k_1$, $\text{rank}(\mathbf{B}) = k_2$, and $\mathbf{AB} = \mathbf{0}$, then $(\mathbf{u}'\mathbf{Au}/k_1)/(\mathbf{u}'\mathbf{Bu}/k_2) \sim F_{k_1,k_2}$.

Summary

This Advanced Treatment contains a condensed form of the background information needed to study the classical linear model using matrices. Although the material here is self-contained, it is primarily intended as a review for readers who are familiar with matrix algebra and multivariate statistics, and it will be used extensively in Advanced Treatment E.

Key Terms

Chi-Square Random Variable
Column Vector
Diagonal Matrix
Expected Value
F Random Variable
Idempotent Matrix
Identity Matrix
Inverse
Linearly Independent Vectors
Matrix
Matrix Multiplication
Multivariate Normal Distribution
Positive Definite (p.d.)
Positive Semi-Definite (p.s.d.)
Quadratic Form

Random Vector
Rank of a Matrix
Row Vector
Scalar Multiplication
Square Matrix
Symmetric Matrix
t Distribution
Trace of a Matrix
Transpose
Variance-Covariance Matrix
Zero Matrix

Problems

1 (i) Find the product **AB** using

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 1 & 6 \\ 1 & 8 & 0 \\ 3 & 0 & 0 \end{bmatrix}.$$

(ii) Does **BA** exist?

2 If **A** and **B** are $n \times n$ diagonal matrices, show that $\mathbf{AB} = \mathbf{BA}$.

3 Let **X** be any $n \times k$ matrix. Show that $\mathbf{X}'\mathbf{X}$ is a symmetric matrix.

4 (i) Use the properties of trace to argue that $\text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}')$ for any $n \times m$ matrix **A**.

(ii) For $\mathbf{A} = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 3 & 0 \end{bmatrix}$, verify that $\text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}')$.

5 (i) Use the definition of inverse to prove the following: if **A** and **B** are $n \times n$ nonsingular matrices, then $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

(ii) If **A**, **B**, and **C** are all $n \times n$ nonsingular matrices, find $(\mathbf{ABC})^{-1}$ in terms of $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$, and $\mathbf{C}^{-1}$.

6 (i) Show that if **A** is an $n \times n$ symmetric, positive semi-definite matrix, then **A** must have nonnegative diagonal elements.

(ii) Show that if **A** is an $n \times n$ symmetric, positive definite matrix, then **A** must have strictly positive diagonal elements.

(iii) Write down a 2×2 symmetric matrix with strictly positive diagonal elements that is *not* positive definite.

7 Let **A** be an $n \times n$ symmetric, positive definite matrix. Show that if **P** is any $n \times n$ nonsingular matrix, then $\mathbf{P}'\mathbf{AP}$ is positive definite.

8 Prove Property 5 of variances for vectors, using Property 3.

9 Let **a** be an $n \times 1$ nonrandom vector and let **u** be an $n \times 1$ random vector with $\text{E}(\mathbf{u}\mathbf{u}') = \mathbf{I}_n$. Show that $\text{E}[\text{tr}(\mathbf{a}\mathbf{u}\mathbf{u}'\mathbf{a}')] = \sum_{i=1}^{n} a_i^2$.

10 Take as given the properties of the chi-square distribution listed in the text. Show how those properties, along with the definition of an *F* random variable, imply the stated property of the *F* distribution (concerning ratios of quadratic forms).

11 Let **X** be an $n \times k$ matrix partitioned as

$$\mathbf{X} = (\mathbf{X}_1 \ \mathbf{X}_2),$$

where $\mathbf{X}_1$ is $n \times k_1$ and $\mathbf{X}_2$ is $n \times k_2$.

(i) Show that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}.$$

What are the dimensions of each of the matrices?

(ii) Let $\mathbf{b}$ be a $k \times 1$ vector, partitioned as

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix},$$

where $\mathbf{b}_1$ is $k_1 \times 1$ and $\mathbf{b}_2$ is $k_2 \times 1$. Show that

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)\mathbf{b}_1 + (\mathbf{X}_1'\mathbf{X}_2)\mathbf{b}_2 \\ (\mathbf{X}_2'\mathbf{X}_1)\mathbf{b}_1 + (\mathbf{X}_2'\mathbf{X}_2)\mathbf{b}_2 \end{pmatrix}.$$

12 (i) Let $\mathbf{A}$ be an $n \times n$ symmetric matrix such that $\mathbf{A}$ and $\mathbf{I}_n - \mathbf{A}$ are both positive semi-definite. Show that $0 \leq a_{ii} \leq 1$ for $i = 1, \ldots, n$, where a_{ii} is the i^{th} diagonal element of $\mathbf{A}$.

(ii) Prove that if $\mathbf{A}$ is an $n \times n$ symmetric, idempotent matrix then it must be positive semi-definite.

(iii) Prove that the only $n \times n$ symmetric, idempotent matrix that is also invertible is $\mathbf{I}_n$.

Advanced Treatment E

The Linear Regression Model in Matrix Form

This Advanced Treatment derives various results for ordinary least squares estimation of the multiple linear regression model using matrix notation and matrix algebra (see Advanced Treatment D for a summary). The material presented here is much more advanced than that in the text.

E-1 The Model and Ordinary Least Squares Estimation

Throughout this Advanced Treatment, we use the t subscript to index observations and an n to denote the sample size. It is useful to write the multiple linear regression model with k parameters as follows:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t, t = 1, 2, \ldots, n, \qquad [E.1]$$

where y_t is the dependent variable for observation t and $x_{tj}, j = 1, 2, \ldots, k$, are the independent variables. As usual, β_0 is the intercept and $\beta_1, \ldots, \beta_k$ denote the slope parameters.

For each t, define a $1 \times (k + 1)$ vector, $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tk})$, and let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ be the $(k + 1) \times 1$ vector of all parameters. Then, we can write (E.1) as

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + u_t, t = 1, 2, \ldots, n. \qquad [E.2]$$

[Some authors prefer to define $\mathbf{x}_t$ as a column vector, in which case $\mathbf{x}_t$ is replaced with $\mathbf{x}_t'$ in (E.2). Mathematically, it makes more sense to define it as a row vector.] We can write (E.2) in full matrix notation by appropriately defining data vectors and matrices. Let $\mathbf{y}$ denote the $n \times 1$ vector of observations on y: the t^{th} element of $\mathbf{y}$ is y_t. Let $\mathbf{X}$ be the $n \times (k + 1)$ vector of observations on the explanatory variables. In other words, the t^{th} row of $\mathbf{X}$ consists of the vector $\mathbf{x}_t$. Written out in detail,

$$\underset{n \times (k+1)}{\mathbf{X}} \equiv \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}.$$

Finally, let $\mathbf{u}$ be the $n \times 1$ vector of unobservable errors or disturbances. Then, we can write (E.2) for all n observations in **matrix notation**:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \qquad \text{[E.3]}$$

Remember, because $\mathbf{X}$ is $n \times (k + 1)$ and $\boldsymbol{\beta}$ is $(k + 1) \times 1$, $\mathbf{X}\boldsymbol{\beta}$ is $n \times 1$.

Estimation of $\boldsymbol{\beta}$ proceeds by minimizing the sum of squared residuals, as in Section 3-2. Define the sum of squared residuals function for any possible $(k + 1) \times 1$ parameter vector $\mathbf{b}$ as

$$\text{SSR}(\mathbf{b}) \equiv \sum_{t=1}^{n} (y_t - \mathbf{x}_t\mathbf{b})^2.$$

The $(k + 1) \times 1$ vector of ordinary least squares estimates, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)'$, minimizes SSR($\mathbf{b}$) over all possible $(k + 1) \times 1$ vectors $\mathbf{b}$. This is a problem in multivariable calculus. For $\hat{\boldsymbol{\beta}}$ to minimize the sum of squared residuals, it must solve the **first order condition**

$$\partial \text{SSR}(\hat{\boldsymbol{\beta}})/\partial \mathbf{b} \equiv 0. \qquad \text{[E.4]}$$

Using the fact that the derivative of $(y_t - \mathbf{x}_t\mathbf{b})^2$ with respect to $\mathbf{b}$ is the $1 \times (k + 1)$ vector $-2(y_t - \mathbf{x}_t\mathbf{b})\mathbf{x}_t$, (E.4) is equivalent to

$$\sum_{t=1}^{n} \mathbf{x}_t'(y_t - \mathbf{x}_t\hat{\boldsymbol{\beta}}) \equiv \mathbf{0}. \qquad \text{[E.5]}$$

(We have divided by −2 and taken the transpose.) We can write this first order condition as

$$\begin{aligned}
&\sum_{t=1}^{n} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \cdots - \hat{\beta}_k x_{tk}) = 0 \\
&\sum_{t=1}^{n} x_{t1}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \cdots - \hat{\beta}_k x_{tk}) = 0 \\
&\quad \vdots \\
&\sum_{t=1}^{n} x_{tk}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \cdots - \hat{\beta}_k x_{tk}) = 0,
\end{aligned}$$

which is identical to the first order conditions in equation (3.13). We want to write these in matrix form to make them easier to manipulate. Using the formula for partitioned multiplication in Advanced Treatment D, we see that (E.5) is equivalent to

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \qquad \text{[E.6]}$$

or

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \qquad \text{[E.7]}$$

It can be shown that (E.7) always has at least one solution. Multiple solutions do not help us, as we are looking for a unique set of OLS estimates given our data set. Assuming that the $(k + 1) \times (k + 1)$ symmetric matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, we can premultiply both sides of (E.7) by $(\mathbf{X}'\mathbf{X})^{-1}$ to solve for the OLS estimator $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \qquad \text{[E.8]}$$

This is the critical formula for matrix analysis of the multiple linear regression model. The assumption that $\mathbf{X}'\mathbf{X}$ is invertible is equivalent to the assumption that rank($\mathbf{X}$) $= (k + 1)$, which means that the columns of $\mathbf{X}$ must be linearly independent. This is the matrix version of MLR.3 in Chapter 3.

Before we continue, (E.8) warrants a word of warning. It is tempting to simplify the formula for $\hat{\boldsymbol{\beta}}$ as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}\mathbf{y}.$$

The flaw in this reasoning is that $\mathbf{X}$ is usually not a square matrix, so it cannot be inverted. In other words, we cannot write $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}$ unless $n = (k + 1)$, a case that virtually never arises in practice.

The $n \times 1$ vectors of OLS fitted values and residuals are given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \text{ respectively.}$$

From (E.6) and the definition of $\hat{\mathbf{u}}$, we can see that the first order condition for $\hat{\boldsymbol{\beta}}$ is the same as

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}. \qquad \text{[E.9]}$$

Because the first column of $\mathbf{X}$ consists entirely of ones, (E.9) implies that the OLS residuals always sum to zero when an intercept is included in the equation and that the sample covariance between each independent variable and the OLS residuals is zero. (We discussed both of these properties in Chapter 3.)

The sum of squared residuals can be written as

$$\text{SSR} = \sum_{t=1}^{n} \hat{u}_t^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \qquad \text{[E.10]}$$

All of the algebraic properties from Chapter 3 can be derived using matrix algebra. For example, we can show that the total sum of squares is equal to the explained sum of squares plus the sum of squared residuals [see (3.27)]. The use of matrices does not provide a simpler proof than summation notation, so we do not provide another derivation.

The matrix approach to multiple regression can be used as the basis for a geometrical interpretation of regression. This involves mathematical concepts that are even more advanced than those we covered in Advanced Treatment D. [See Goldberger (1991) or Greene (1997).]

E-1a The Frisch-Waugh Theorem

In Section 3-2, we described a "partialling out" interpretation of the ordinary least squares estimates. We can establish the partialling out interpretation very generally using matrix notation. Partition the $n \times (k + 1)$ matrix $\mathbf{X}$ as

$$\mathbf{X} = (\mathbf{X}_1|\mathbf{X}_2),$$

where $\mathbf{X}_1$ is $n \times (k_1 + 1)$ and includes the intercept—although that is not required for the result to hold—and $\mathbf{X}_2$ is $n \times k_2$. We still assume that $\mathbf{X}$ has rank $k + 1$, which means $\mathbf{X}_1$ has rank $k_1 + 1$ and $\mathbf{X}_2$ has rank k_2.

Consider the OLS estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ from the (long) regression

$$\mathbf{y} \text{ on } \mathbf{X}_1, \mathbf{X}_2.$$

As we know, the multiple regression coefficients on $\mathbf{X}_2$, $\hat{\boldsymbol{\beta}}_2$, generally differs from $\tilde{\boldsymbol{\beta}}_2$ from the regression $\mathbf{y}$ on $\mathbf{X}_2$. One way to describe the difference is to understand that we can obtain $\hat{\boldsymbol{\beta}}_2$ from a shorter regression, but first we must "partial out" $\mathbf{X}_1$ from $\mathbf{X}_2$. Consider the following two-step method:

(i) Regress (each column of) $\mathbf{X}_2$ on $\mathbf{X}_1$ and obtain the matrix of residuals, say $\ddot{\mathbf{X}}_2$. We can write $\ddot{\mathbf{X}}_2$ as

$$\ddot{\mathbf{X}}_2 = [\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}_2 = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 = \mathbf{M}_1\mathbf{X}_2,$$

where $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ and $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_1$ are $n \times n$ symmetric, idempotent matrices.

(ii) Regress $\mathbf{y}$ on $\ddot{\mathbf{X}}_2$ and call the $k_2 \times 1$ vector of coefficient $\ddot{\boldsymbol{\beta}}_2$.

The **Frisch-Waugh (FW) theorem** states that

$$\ddot{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_2.$$

Importantly, the FW theorem generally says nothing about equality of the estimates from the long regression, $\hat{\boldsymbol{\beta}}_2$, and those from the short regression, $\tilde{\boldsymbol{\beta}}_2$. Usually $\hat{\boldsymbol{\beta}}_2 \neq \tilde{\boldsymbol{\beta}}_2$. However, if $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ then $\ddot{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{X}_2 = \mathbf{X}_2$, in which case $\ddot{\boldsymbol{\beta}}_2 = \tilde{\boldsymbol{\beta}}_2$; then $\hat{\boldsymbol{\beta}}_2 = \tilde{\boldsymbol{\beta}}_2$ follows from FW. It is also worth noting that we obtain $\hat{\boldsymbol{\beta}}_2$ if we also partial $\mathbf{X}_1$ out of $\mathbf{y}$. In other words, let $\ddot{\mathbf{y}}$ be the residuals from regressing $\mathbf{y}$ on $\mathbf{X}_1$, so that

$$\ddot{\mathbf{y}} = \mathbf{M}_1\mathbf{y}.$$

Then $\hat{\boldsymbol{\beta}}_2$ is obtained from the regression $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{X}}_2$. It is important to understand that it is not enough to only partial out $\mathbf{X}_1$ from $\mathbf{y}$. The important step is partialling out $\mathbf{X}_1$ from $\mathbf{X}_2$. Problem 6 at the end of this chapter asks you to derive the FW theorem and to investigate some related issues.

Another useful algebraic result is that when we regress $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{X}}_2$ and save the residuals, say $\ddot{\mathbf{u}}$, these are identical to the OLS residuals from the original (long) regression:

$$\ddot{\mathbf{y}} = \ddot{\mathbf{X}}_2\hat{\boldsymbol{\beta}}_2 = \ddot{\mathbf{u}} = \hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2,$$

where we have used the FW result $\ddot{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_2$. We do not obtain the original OLS residuals if we regress $\mathbf{y}$ on $\ddot{\mathbf{X}}_2$ (but we do obtain $\hat{\boldsymbol{\beta}}_2$).

Before the advent of powerful computers, the Frisch-Waugh result was sometimes used as a computational device. Today, the result is more of theoretical interest, and it is very helpful in understanding the mechanics of OLS. For example, recall that in Chapter 10 we used the FW theorem to establish that adding a time trend to a multiple regression is algebraically equivalent to first linearly detrending all of the explanatory variables before running the regression. The FW theorem also can be used in Chapter 14 to establish that the fixed effects estimator, which we introduced as being obtained from OLS on time-demeaned data, can also be obtained from the (long) dummy variable regression.

E-2 Finite Sample Properties of OLS

Deriving the expected value and variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ is facilitated by matrix algebra, but we must show some care in stating the assumptions.

Assumption E.1 **Linear in Parameters**

The model can be written as in (E.3), where $\mathbf{y}$ is an observed $n \times 1$ vector, $\mathbf{X}$ is an $n \times (k + 1)$ observed matrix, and $\mathbf{u}$ is an $n \times 1$ vector of unobserved errors or disturbances.

Assumption E.2 **No Perfect Collinearity**

The matrix $\mathbf{X}$ has rank $(k + 1)$.

This is a careful statement of the assumption that rules out linear dependencies among the explanatory variables. Under Assumption E.2, $\mathbf{X}'\mathbf{X}$ is nonsingular, so $\hat{\boldsymbol{\beta}}$ is unique and can be written as in (E.8).

Assumption E.3 **Zero Conditional Mean**

Conditional on the entire matrix $\mathbf{X}$, each error u_t has zero mean: $\mathrm{E}(u_t|\mathbf{X}) = 0$, $t = 1, 2, \ldots, n$.

In vector form, Assumption E.3 can be written as

$$E(\mathbf{u}|\mathbf{X}) = \mathbf{0}. \qquad [E.11]$$

This assumption is implied by MLR.4 under the random sampling assumption, MLR.2. In time series applications, Assumption E.3 imposes strict exogeneity on the explanatory variables, something discussed at length in Chapter 10. This rules out explanatory variables whose future values are correlated with u_t; in particular, it eliminates lagged dependent variables. Under Assumption E.3, we can condition on the x_{tj} when we compute the expected value of $\hat{\boldsymbol{\beta}}$.

THEOREM E.1

UNBIASEDNESS OF OLS

Under Assumptions E.1, E.2, and E.3, the OLS estimator $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

***PROOF*:** Use Assumptions E.1 and E.2 and simple algebra to write

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}, \qquad [E.12]\end{aligned}$$

where we use the fact that $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}_{k+1}$. Taking the expectation conditional on $\mathbf{X}$ gives

$$\begin{aligned}E(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0} = \boldsymbol{\beta},\end{aligned}$$

because $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ under Assumption E.3. This argument clearly does not depend on the value of $\boldsymbol{\beta}$, so we have shown that $\hat{\boldsymbol{\beta}}$ is unbiased.

To obtain the simplest form of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, we impose the assumptions of homoskedasticity and no serial correlation.

Assumption E.4 (Homoskedasticity)

Conditional on $\mathbf{X}$, the variances are constant:

$$\text{Var}(u_t|\mathbf{X}) = \sigma^2, t = 1, \ldots, n. \square$$

As we discussed throughout the text, especially in Chapters 8 and 12, heteroskedasticity—which is failure of E.4—can never be ruled out for any of the data structures (cross section, time series, panel).

Assumption E.5 (No Serial Correlation)

Conditional on $\mathbf{X}$, the errors are uncorrelated for all $t \neq s$:

$$\text{Cov}(u_t,u_s|\mathbf{X}) = 0, \text{ all } t \neq s. \square$$

Assumption E.5 is automatically satisfied under random sampling, which is why it does not appear until Chapter 10. With time series applications, Assumption E.5 means that the errors or innovations are uncorrelated across time. As we discussed in Chapters 10, 11, and 12, Assumption E.5 can be unrealistic, particularly in models that do not include lags of y_t. (Including, say, y_{t-1} in $\mathbf{x}_t$ is ruled out by Assumption E.3.)

We can combine Assumptions E.4 and E.5 into a simple expression using matrix notation:

$$\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2 \mathbf{I}_n. \qquad [E.13]$$

Under this assumption, the $n \times n$ variance-covariance matrix $\text{Var}(\mathbf{u}|\mathbf{X})$ depends only on a single parameter, σ^2, and we often say that $\mathbf{u}$ has a **scalar variance-covariance matrix**. (The "scalar" is σ^2.)

Assumptions E.1 through E.5 comprise the **Gauss-Markov assumptions.** The statements of the assumptions unify the conditions we used for cross-sectional analysis in Chapter 3 and time series analysis in Chapter 10.

Using the concise expression in (E.13), we can derive the **variance-covariance matrix of the OLS estimator** under the Gauss-Markov Assumptions.

THEOREM E.2

VARIANCE-COVARIANCE MATRIX OF THE OLS ESTIMATOR

Under Assumptions E.1 through E.5,

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad \text{[E.14]}$$

PROOF: From the last formula in equation (E.12), we have

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{Var}(\mathbf{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Now, we use equation (E.13) to get

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Expression (E.14) means that the variance of $\hat{\beta}_j$ (conditional on $\mathbf{X}$) is obtained by multiplying σ^2 by the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. For the slope coefficients, we gave an interpretable formula in equation (3.51). Equation (E.14) also tells us how to obtain the covariance between any two OLS estimates: multiply σ^2 by the appropriate off-diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. In Chapter 4, we showed how to avoid explicitly finding covariances for obtaining confidence intervals and hypothesis tests by appropriately rewriting the model.

The Gauss-Markov Theorem, in its full generality, can be proven.

THEOREM E.3

GAUSS-MARKOV THEOREM

Under Assumptions E.1 through E.5, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator.

PROOF: Any other linear estimator of $\boldsymbol{\beta}$ can be written as

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y}, \qquad \text{[E.15]}$$

where $\mathbf{A}$ is an $n \times (k+1)$ matrix. In order for $\tilde{\boldsymbol{\beta}}$ to be unbiased conditional on $\mathbf{X}$, $\mathbf{A}$ can consist of nonrandom numbers and functions of $\mathbf{X}$. (For example, $\mathbf{A}$ cannot be a function of $\mathbf{y}$.) To see what further restrictions on $\mathbf{A}$ are needed, write

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = (\mathbf{A}'\mathbf{X})\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}. \qquad \text{[E.16]}$$

Then,

$$\begin{aligned}\text{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \text{E}(\mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\text{E}(\mathbf{u}|\mathbf{X}) \text{ because } \mathbf{A} \text{ is a function of } \mathbf{X} \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} \text{ because } \text{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}.\end{aligned}$$

For $\tilde{\boldsymbol{\beta}}$ to be an unbiased estimator of $\boldsymbol{\beta}$, it must be true that $E(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$ for all $(k+1) \times 1$ vectors $\boldsymbol{\beta}$, that is,

$$\mathbf{A}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \text{ for all } (k+1) \times 1 \text{ vectors } \boldsymbol{\beta}. \qquad [E.17]$$

Because $\mathbf{A}'\mathbf{X}$ is a $(k+1) \times (k+1)$ matrix, (E.17) holds if, and only if, $\mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1}$. Equations (E.15) and (E.17) characterize the class of linear, unbiased estimators for $\boldsymbol{\beta}$.

Next, from (E.16), we have

$$\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{A}'[\text{Var}(\mathbf{u}|\mathbf{X})]\mathbf{A} = \sigma^2\mathbf{A}'\mathbf{A},$$

by equation (E.13). Therefore,

$$\begin{aligned}\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2[\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2[\mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}] \text{ because } \mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1} \\ &= \sigma^2\mathbf{A}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{A} \\ &\equiv \sigma^2\mathbf{A}'\mathbf{M}\mathbf{A},\end{aligned}$$

where $\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Because $\mathbf{M}$ is symmetric and idempotent, $\mathbf{A}'\mathbf{M}\mathbf{A}$ is positive semi-definite for any $n \times (k+1)$ matrix $\mathbf{A}$. This establishes that the OLS estimator $\hat{\boldsymbol{\beta}}$ is BLUE. Why is this important? Let $\mathbf{c}$ be any $(k+1) \times 1$ vector and consider the linear combination $\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \cdots + c_k\beta_k$, which is a scalar. The unbiased estimators of $\mathbf{c}'\boldsymbol{\beta}$ are $\mathbf{c}'\hat{\boldsymbol{\beta}}$ and $\mathbf{c}'\tilde{\boldsymbol{\beta}}$. But

$$\text{Var}(\mathbf{c}'\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{c}'[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X})]\mathbf{c} \geq 0,$$

because $[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X})]$ is p.s.d. Therefore, when it is used for estimating any linear combination of $\boldsymbol{\beta}$, OLS yields the smallest variance. In particular, $\text{Var}(\hat{\beta}_j|\mathbf{X}) \leq \text{Var}(\tilde{\beta}_j\mathbf{X})$ for any other linear, unbiased estimator of β_j.

The unbiased estimator of the error variance σ^2 can be written as

$$\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(n-k-1),$$

which is the same as equation (3.56).

THEOREM E.4

UNBIASEDNESS OF $\hat{\sigma}^2$

Under Assumptions E.1 through E.5, $\hat{\sigma}^2$ is unbiased for σ^2: $E(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ for all $\sigma^2 > 0$.

PROOF: Write $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{u}$, where $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and the last equality follows because $\mathbf{M}\mathbf{X} = \mathbf{0}$. Because $\mathbf{M}$ is symmetric and idempotent,

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u}.$$

Because $\mathbf{u}'\mathbf{M}\mathbf{u}$ is a scalar, it equals its trace. Therefore,

$$\begin{aligned}E(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X}) &= E[\text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u})|\mathbf{X}] = E[\text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')|\mathbf{X}] \\ &= \text{tr}[E(\mathbf{M}\mathbf{u}\mathbf{u}'|\mathbf{X})] = \text{tr}[\mathbf{M}E(\mathbf{u}\mathbf{u}'|\mathbf{X})] \\ &= \text{tr}(\mathbf{M}\sigma^2\mathbf{I}_n) = \sigma^2\text{tr}(\mathbf{M}) = \sigma^2(n-k-1).\end{aligned}$$

The last equality follows from $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = n - \text{tr}(\mathbf{I}_{k+1}) = n - (k+1) = n - k - 1$. Therefore,

$$E(\hat{\sigma}^2|\mathbf{X}) = E(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X})/(n-k-1) = \sigma^2.$$

E-3 Statistical Inference

When we add the final classical linear model assumption, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution, which leads to the t and F distributions for the standard test statistics covered in Chapter 4.

Assumption E.6 **Normality of Errors**

Conditional on $\mathbf{X}$, the u_t are independent and identically distributed as Normal$(0,\sigma^2)$. Equivalently, $\mathbf{u}$ given $\mathbf{X}$ is distributed as multivariate normal with mean zero and variance-covariance matrix $\sigma^2\mathbf{I}_n$: $u \sim \text{Normal}(0,\sigma^2\mathbf{I}_n)$.

Assumption E.6 implies Assumptions E.3, E.4, and E.5, but it is much stronger because it assumes that each u_t has a Normal$(0, \sigma^2)$ distribution. As a technical point, Assumption E.6 implies that the u_t are actually independent across t rather than merely uncorrelated. From a practical perspective, this distinction is unimportant. Assumptions E.1 through E.6 are the **classical linear model (CLM) assumptions** expressed in matrix terms, and they are usually viewed as the Gauss-Markov assumptions plus normality of the errors.

THEOREM E.5 **NORMALITY OF $\hat{\beta}$**

Under the classical linear model Assumptions E.1 through E.6, $\hat{\boldsymbol{\beta}}$ conditional on $\mathbf{X}$ is distributed as multivariate normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Theorem E.5 is the basis for statistical inference involving $\boldsymbol{\beta}$. In fact, along with the properties of the chi-square, t, and F distributions that we summarized in Advanced Treatment D, we can use Theorem E.5 to establish that t statistics have a t distribution under Assumptions E.1 through E.6 (under the null hypothesis) and likewise for F statistics. We illustrate with a proof for the t statistics.

THEOREM E.6 **DISTRIBUTION OF *t* STATISTIC**

Under Assumptions E.1 through E.6,

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \sim t_{n-k-1}, j = 0, 1, \ldots, k.$$

***PROOF*:** The proof requires several steps; the following statements are initially conditional on $\mathbf{X}$. First, by Theorem E.5, $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \sim \text{Normal}(0,1)$, where $\text{sd}(\hat{\beta}_j) = \sigma\sqrt{C_{jj}}$, and c_{jj} is the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Next, under Assumptions E.1 through E.6, conditional on $\mathbf{X}$,

$$(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}. \qquad \text{[E.18]}$$

This follows because $(n - k - 1)\hat{\sigma}^2/\sigma^2 = (\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma)$, where $\mathbf{M}$ is the $n \times n$ symmetric, idempotent matrix defined in Theorem E.3. But $\mathbf{u}/\sigma \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ by Assumption E.6. It follows from Property 1 for the chi-square distribution in Advanced Treatment D that $(\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma) \sim \chi^2_{n-k-1}$ (because $\mathbf{M}$ has rank $n - k - 1$).

We also need to show that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. But $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$, and $\hat{\sigma}^2 = \mathbf{u}'\mathbf{M}\mathbf{u}/(n - k - 1)$. Now, $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{M} = \mathbf{0}$ because $\mathbf{X}'\mathbf{M} = \mathbf{0}$. It follows, from Property 5 of the

multivariate normal distribution in Advanced Treatment D, that $\hat{\boldsymbol{\beta}}$ and **Mu** are independent. Because $\hat{\sigma}^2$ is a function of **Mu**, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are also independent.

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) = [(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j)]/(\hat{\sigma}^2/\sigma^2)^{1/2},$$

which is the ratio of a standard normal random variable and the square root of a $\chi^2_{n-k-1}/(n-k-1)$ random variable. We just showed that these are independent, so, by definition of a t random variable, $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ has the t_{n-k-1} distribution. Because this distribution does not depend on **X**, it is the unconditional distribution of $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ as well.

From this theorem, we can plug in any hypothesized value for β_j and use the t statistic for testing hypotheses, as usual.

Under Assumptions E.1 through E.6, we can compute what is known as the *Cramer-Rao* lower bound for the variance-covariance matrix of unbiased estimators of $\boldsymbol{\beta}$ (again conditional on **X**) [see Greene (1997, Chapter 4)]. This can be shown to be $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, which is exactly the variance-covariance matrix of the OLS estimator. This implies that $\hat{\boldsymbol{\beta}}$ is the **minimum variance unbiased estimator** of $\boldsymbol{\beta}$ (conditional on **X**): $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ is positive semi-definite for any other unbiased estimator $\tilde{\boldsymbol{\beta}}$; we no longer have to restrict our attention to estimators linear in **y**.

It is easy to show that the OLS estimator is in fact the maximum likelihood estimator of $\boldsymbol{\beta}$ under Assumption E.6. For each t, the distribution of y_t given **X** is $\text{Normal}(\mathbf{x}_t\boldsymbol{\beta},\sigma^2)$. Because the y_t are independent conditional on **X**, the likelihood function for the sample is obtained from the product of the densities:

$$\prod_{t=1}^{n} (2\pi\sigma^2)^{-1/2}\exp[-(y_t - \mathbf{x}_t\boldsymbol{\beta})^2/(2\sigma^2)],$$

where Π denotes product. Maximizing this function with respect to $\boldsymbol{\beta}$ and σ^2 is the same as maximizing its natural logarithm:

$$\sum_{t=1}^{n} [-(1/2)\log(2\pi\sigma^2) - (y_t - \mathbf{x}_t\boldsymbol{\beta})^2/(2\sigma^2)].$$

For obtaining $\hat{\boldsymbol{\beta}}$, this is the same as minimizing $\sum_{t=1}^{n}(y_t - \mathbf{x}_t\boldsymbol{\beta})^2$—the division by $2\sigma^2$ does not affect the optimization—which is just the problem that OLS solves. The estimator of σ^2 that we have used, $\text{SSR}/(n-k)$, turns out not to be the MLE of σ^2; the MLE is SSR/n, which is a biased estimator. Because the unbiased estimator of σ^2 results in t and F statistics with exact t and F distributions under the null, it is always used instead of the MLE.

That the OLS estimator is the MLE under Assumption E.6 implies an interesting robustness property of the MLE based on the normal distribution. The reasoning is simple. We know that the OLS estimator is unbiased under Assumptions E.1 to E.3; normality of the errors is used nowhere in the proof, and neither are Assumptions E.4 and E.5. As the next section shows, the OLS estimator is also consistent without normality, provided the law of large numbers holds (as is widely true). These statistical properties of the OLS estimator imply that the MLE based on the normal log-likelihood function is robust to distributional specification: the distribution can be (almost) anything and yet we still obtain a consistent (and, under E.1 to E.3, unbiased) estimator. As discussed in Section 17-3, a maximum likelihood estimator obtained without assuming the distribution is correct is often called a **quasi-maximum likelihood estimator (QMLE).**

Generally, consistency of the MLE relies on having a correct distribution in order to conclude that it is consistent for the parameters. We have just seen that the normal distribution is a notable exception. There are some other distributions that share this property, including the Poisson distribution—as discussed in Section 17-3. Wooldridge (2010, Chapter 18) discusses some other useful examples.

E-4 Some Asymptotic Analysis

The matrix approach to the multiple regression model can also make derivations of asymptotic properties more concise. In fact, we can give general proofs of the claims in Chapter 11.

We begin by proving the consistency result of Theorem 11.1. Recall that these assumptions contain, as a special case, the assumptions for cross-sectional analysis under random sampling.

Proof of Theorem 11.1. As in Problem E.1 and using Assumption TS.1′ we write the OLS estimator as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'y_t\right) = \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'(\mathbf{x}_t\boldsymbol{\beta} + u_t)\right) \\ &= \boldsymbol{\beta} + \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) \qquad \text{[E.19]} \\ &= \boldsymbol{\beta} + \left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right).\end{aligned}$$

Now, by the law of large numbers,

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t \xrightarrow{p} \mathbf{A} \text{ and } n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t \xrightarrow{p} \mathbf{0}, \qquad \text{[E.20]}$$

where $\mathbf{A} = \mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$ is a $(k+1)\times(k+1)$ nonsingular matrix under Assumption TS.2′ and we have used the fact that $\mathrm{E}(\mathbf{x}_t'u_t) = 0$ under Assumption TS.3′. Now, we must use a matrix version of Property PLIM.1 in Math Refresher C. Namely, because $\mathbf{A}$ is nonsingular,

$$\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1} \xrightarrow{p} \mathbf{A}^{-1}. \qquad \text{[E.21]}$$

[Wooldridge (2010, Chapter 3) contains a discussion of these kinds of convergence results.] It now follows from (E.19), (E.20), and (E.21) that

$$\operatorname{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{A}^{-1}\cdot\mathbf{0} = \boldsymbol{\beta}.$$

This completes the proof.

Next, we sketch a proof of the asymptotic normality result in Theorem 11.2.

Proof of Theorem 11.2. From equation (E.19), we can write

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) \\ &= \mathbf{A}^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) + o_p(1), \qquad \text{[E.22]}\end{aligned}$$

where the term "$o_p(1)$" is a remainder term that converges in probability to zero. This term is equal to $[(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t)^{-1} - \mathbf{A}^{-1}](n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t)$. The term in brackets converges in probability to zero (by the same argument used in the proof of Theorem 11.1), while $(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t)$ is bounded in probability because it converges to a multivariate normal distribution by the central limit theorem. A well-known result in asymptotic theory is that the product of such terms converges in probability to zero. Further, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ inherits its asymptotic distribution from $\mathbf{A}^{-1}(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t)$. See Wooldridge (2010, Chapter 3) for more details on the convergence results used in this proof.

By the central limit theorem, $n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t$ has an asymptotic normal distribution with mean zero and, say, $(k + 1) \times (k + 1)$ variance-covariance matrix $\mathbf{B}$. Then, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has an asymptotic multivariate normal distribution with mean zero and variance-covariance matrix $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. We now show that, under Assumptions TS.4′ and TS.5′, $\mathbf{B} = \sigma^2\mathbf{A}$. (The general expression is useful because it underlies heteroskedasticity-robust and serial correlation-robust standard errors for OLS, of the kind discussed in Chapter 12.) First, under Assumption TS.5′ $\mathbf{x}_t'u_t$ and $\mathbf{x}_s'u_s$ are uncorrelated for $t \neq s$. Why? Suppose $s < t$ for concreteness. Then, by the law of iterated expectations, $\mathrm{E}(\mathbf{x}_t'u_tu_s\mathbf{x}_s) = \mathrm{E}[\mathrm{E}(u_tu_s\mathbf{x}_t'\mathbf{x}_s)|\mathbf{x}_t,\mathbf{x}_s] = \mathrm{E}[0 \cdot \mathbf{x}_t'\mathbf{x}_s] = 0$. The zero covariances imply that the variance of the sum is the sum of the variances. But $\mathrm{Var}(\mathbf{x}_t'u_t) = \mathrm{E}(\mathbf{x}_t'u_tu_t\mathbf{x}_t) = \mathrm{E}(u_y^2\mathbf{x}_t'\mathbf{x}_t)$. By the law of iterated expectations, $\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t) = \mathrm{E}[\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t|\mathbf{x}_t) = \mathrm{E}[\mathrm{E}(u_t^2|\mathbf{x}_t)\mathbf{x}_t'\mathbf{x}_t] = \mathrm{E}[\sigma^2\mathbf{x}_t'\mathbf{x}_t] = \sigma^2\mathrm{E}(\mathbf{x}_t'\mathbf{x}_t) = \sigma^2\mathbf{A}$, where we use $\mathrm{E}(u_t^2|\mathbf{x}_t) = \sigma^2$ under Assumptions TS.3′ and TS.4′. This shows that $\mathbf{B} = \sigma^2\mathbf{A}$, and so, under Assumptions TS.1′ to TS.5′, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} \mathrm{Normal}(\mathbf{0},\sigma^2\mathbf{A}^{-1}). \qquad \text{[E.23]}$$

This completes the proof.

From equation (E.23), we treat $\hat{\boldsymbol{\beta}}$ as if it is approximately normally distributed with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2\mathrm{A}^{-1}/n$. The division by the sample size, n, is expected here: the approximation to the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ shrinks to zero at the rate $1/n$. When we replace σ^2 with its consistent estimator, $\hat{\sigma}^2 = \mathrm{SSR}/(n - k - 1)$, and replace A with its consistent estimator, $n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t = \mathbf{X}'\mathbf{X}/n$, we obtain an estimator for the asymptotic variance of $\hat{\boldsymbol{\beta}}$:

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad \text{[E.24]}$$

Notice how the two divisions by n cancel, and the right-hand side of (E.24) is just the usual way we estimate the variance matrix of the OLS estimator under the Gauss-Markov assumptions. To summarize, we have shown that, under Assumptions TS.1′ to TS.5′—which contain MLR.1 to MLR.5 as special cases—the usual standard errors and t statistics are asymptotically valid. It is perfectly legitimate to use the usual t distribution to obtain critical values and p-values for testing a single hypothesis. Interestingly, in the general setup of Chapter 11, assuming normality of the errors—say, u_t given $\mathbf{x}_t$, u_{t-1}, $\mathbf{x}_{t-1}$, . . . , u_1, $\mathbf{x}_1$ is distributed as Normal$(0, \sigma^2)$—does not necessarily help, as the t statistics would not generally have exact t statistics under this kind of normality assumption. When we do not assume strict exogeneity of the explanatory variables, exact distributional results are difficult, if not impossible, to obtain.

If we modify the argument above, we can derive a heteroskedasticity-robust, variance-covariance matrix. The key is that we must estimate $\mathrm{E}(u_t^2\mathbf{x}_t'\mathbf{x}_t)$ separately because this matrix no longer equals $\sigma^2\mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$. But, if the $\hat{u}_t$ are the OLS residuals, a consistent estimator is

$$(n - k - 1)^{-1}\sum_{t=1}^{n}\hat{u}_t^2\mathbf{x}_t'\mathbf{x}_t, \qquad \text{[E.25]}$$

where the division by $n - k - 1$ rather than n is a degrees of freedom adjustment that typically helps the finite sample properties of the estimator. When we use the expression in equation (E.25), we obtain

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}) = [n/(n - k - 1)](\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{t=1}^{n}\hat{u}_t^2\mathbf{x}_t'\mathbf{x}_t\right)(\mathbf{X}'\mathbf{X})^{-1}. \qquad \text{[E.26]}$$

The square roots of the diagonal elements of this matrix are the same heteroskedasticity-robust standard errors we obtained in Section 8-2 for the pure cross-sectional case. A matrix extension of the serial correlation- (and heteroskedasticity-) robust standard errors we obtained in Section 12-5 is also available, but the matrix that must replace (E.25) is complicated because of the serial correlation. See, for example, Hamilton (1994, Section 10-5).

E-4a Wald Statistics for Testing Multiple Hypotheses

Similar arguments can be used to obtain the asymptotic distribution of the **Wald statistic** for testing multiple hypotheses. Let $\mathbf{R}$ be a $q \times (k+1)$ matrix, with $q \leq (k+1)$. Assume that the q restrictions on the $(k+1) \times 1$ vector of parameters, $\boldsymbol{\beta}$, can be expressed as H_0: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, where $\mathbf{r}$ is a $q \times 1$ vector of known constants. Under Assumptions TS.1′ to TS.5′, it can be shown that, under H_0,

$$[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})]'(\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})] \overset{a}{\sim} \chi^2_q, \qquad \text{[E.27]}$$

where $\mathbf{A} = \mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$, as in the proofs of Theorems 11.1 and 11.2. The intuition behind equation (E.25) is simple. Because $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is roughly distributed as Normal$(\mathbf{0},\sigma^2\mathbf{A}^{-1})$, $\mathbf{R}[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sqrt{n}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is approximately Normal$(0,\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ by Property 3 of the multivariate normal distribution in Advanced Treatment D. Under H_0, $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, so $\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \overset{a}{\sim}$ Normal$(\mathbf{0},\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ under H_0. By Property 3 of the chi-square distribution, $z'(\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}z \sim \chi^2_q$ if $z \sim$ Normal$(\mathbf{0},\sigma^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$. To obtain the final result formally, we need to use an asymptotic version of this property, which can be found in Wooldridge (2010, Chapter 3).

Given the result in (E.25), we obtain a computable statistic by replacing $\mathbf{A}$ and σ^2 with their consistent estimators; doing so does not change the asymptotic distribution. The result is the so-called Wald statistic, which, after canceling the sample sizes and doing a little algebra, can be written as

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/\hat{\sigma}^2. \qquad \text{[E.28]}$$

Under H_0, $W \overset{a}{\sim} \chi^2_q$, where we recall that q is the number of restrictions being tested. If $\hat{\sigma}^2 = \mathrm{SSR}/(n-k-1)$, it can be shown that W/q is exactly the F statistic we obtained in Chapter 4 for testing multiple linear restrictions. [See, for example, Greene (1997, Chapter 7).] Therefore, under the classical linear model assumptions TS.1 to TS.6 in Chapter 10, W/q has an exact $F_{q,n-k-1}$ distribution. Under Assumptions TS.1′ to TS.5′, we only have the asymptotic result in (E.26). Nevertheless, it is appropriate, and common, to treat the usual F statistic as having an approximate $F_{q,n-k-1}$ distribution.

A Wald statistic that is robust to heteroskedasticity of unknown form is obtained by using the matrix in (E.26) in place of $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, and similarly for a test statistic robust to both heteroskedasticity and serial correlation. The robust versions of the test statistics cannot be computed via sums of squared residuals or R-squareds from the restricted and unrestricted regressions.

Summary

This Advanced Treatment has provided a brief treatment of the linear regression model using matrix notation. This material is included for more advanced classes that use matrix algebra, but it is not needed to read the text. In effect, this Advanced Treatment proves some of the results that we either stated without proof, proved only in special cases, or proved through a more cumbersome method of proof. Other topics—such as asymptotic properties, instrumental variables estimation, and panel data models—can be given concise treatments using matrices. Advanced texts in econometrics, including Davidson and MacKinnon (1993), Greene (1997), Hayashi (2000), and Wooldridge (2010), can be consulted for details.

Key Terms

Classical Linear Model (CLM) Assumptions
First Order Condition
Frisch-Waugh (FW) theorem
Gauss-Markov Assumptions
Matrix Notation
Minimum Variance Unbiased Estimator
Scalar Variance-Covariance Matrix
Variance-Covariance Matrix of the OLS Estimator
Wald Statistic
Quasi-Maximum Likelihood Estimator (QMLE)

Problems

1 Let $\mathbf{x}_t$ be the $1 \times (k + 1)$ vector of explanatory variables for observation t. Show that the OLS estimator $\hat{\boldsymbol{\beta}}$ can be written as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{n} \mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n} \mathbf{x}_t'y_t\right).$$

Dividing each summation by n shows that $\hat{\boldsymbol{\beta}}$ is a function of sample averages.

2 Let $\hat{\boldsymbol{\beta}}$ be the $(k + 1) \times 1$ vector of OLS estimates.

(i) Show that for any $(k + 1) \times 1$ vector $\mathbf{b}$, we can write the sum of squared residuals as

$$\text{SSR}(\mathbf{b}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} + (\hat{\boldsymbol{\beta}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b}).$$

{*Hint*: Write $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = [\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]'[\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]$ and use the fact that $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$.}

(ii) Explain how the expression for SSR($\mathbf{b}$) in part (i) proves that $\hat{\boldsymbol{\beta}}$ uniquely minimizes SSR($\mathbf{b}$) over all possible values of $\mathbf{b}$, assuming $\mathbf{X}$ has rank $k + 1$.

3 Let $\hat{\boldsymbol{\beta}}$ be the OLS estimate from the regression of $\mathbf{y}$ on $\mathbf{X}$. Let $\mathbf{A}$ be a $(k + 1) \times (k + 1)$ nonsingular matrix and define $\mathbf{z}_t \equiv \mathbf{x}_t\mathbf{A}$, $t = 1, \ldots, n$. Therefore, $\mathbf{z}_t$ is $1 \times (k + 1)$ and is a nonsingular linear combination of $\mathbf{x}_t$. Let $\mathbf{Z}$ be the $n \times (k + 1)$ matrix with rows $\mathbf{z}_t$. Let $\tilde{\boldsymbol{\beta}}$ denote the OLS estimate from a regression of $\mathbf{y}$ on $\mathbf{Z}$.

(i) Show that $\tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}$.

(ii) Let $\hat{y}_t$ be the fitted values from the original regression and let $\tilde{y}_t$ be the fitted values from regressing $\mathbf{y}$ on $\mathbf{Z}$. Show that $\tilde{y}_t = \hat{y}_t$, for all $t = 1, 2, \ldots, n$. How do the residuals from the two regressions compare?

(iii) Show that the estimated variance matrix for $\tilde{\boldsymbol{\beta}}$ is $\hat{\sigma}^2\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}^{-1'}$, where $\hat{\sigma}^2$ is the usual variance estimate from regressing $\mathbf{y}$ on $\mathbf{X}$.

(iv) Let the $\hat{\beta}_j$ be the OLS estimates from regressing y_t on $1, x_{t1}, \ldots, x_{tk}$, and let the $\tilde{\beta}_j$ be the OLS estimates from the regression of y_t on $1, a_1x_{t1}, \ldots, a_kx_{tk}$, where $a_i \neq 0$, $j = 1, \ldots, k$. Use the results from part (i) to find the relationship between the $\tilde{\beta}_j$ and the $\hat{\beta}_j$.

(v) Assuming the setup of part (iv), use part (iii) to show that $\text{se}(\tilde{\beta}_j) = \text{se}(\hat{\beta}_j)/|a_j|$.

(vi) Assuming the setup of part (iv), show that the absolute values of the t statistics for $\tilde{\beta}_j$ and $\hat{\beta}_j$ are identical.

4 Assume that the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfies the Gauss-Markov assumptions, let $\mathbf{G}$ be a $(k + 1) \times (k + 1)$ nonsingular, nonrandom matrix, and define $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$, so that $\boldsymbol{\delta}$ is also a $(k + 1) \times 1$ vector. Let $\hat{\boldsymbol{\beta}}$ be the $(k + 1) \times 1$ vector of OLS estimators and define $\hat{\boldsymbol{\delta}} = \mathbf{G}\hat{\boldsymbol{\beta}}$ as the OLS estimator of $\boldsymbol{\delta}$.

(i) Show that $\text{E}(\hat{\boldsymbol{\delta}}|\mathbf{X}) = \boldsymbol{\delta}$.

(ii) Find $\text{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ in terms of σ^2, $\mathbf{X}$, and $\mathbf{G}$.

(iii) Use Problem E.3 to verify that $\hat{\boldsymbol{\delta}}$ and the appropriate estimate of $\text{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ are obtained from the regression of $\mathbf{y}$ on $\mathbf{X}\mathbf{G}^{-1}$.

(iv) Now, let $\mathbf{c}$ be a $(k + 1) \times 1$ vector with at least one nonzero entry. For concreteness, assume that $c_k \neq 0$. Define $\theta = \mathbf{c}'\boldsymbol{\beta}$, so that θ is a scalar. Define $\delta_j = \boldsymbol{\beta}_j$, $j = 0, 1, \ldots, k - 1$ and $\delta_k = \theta$. Show how to define a $(k + 1) \times (k + 1)$ nonsingular matrix $\mathbf{G}$ so that $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$. (*Hint*: Each of the first k rows of $\mathbf{G}$ should contain k zeros and a one. What is the last row?)

(v) Show that for the choice of $\mathbf{G}$ in part (iv),

$$\mathbf{G}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & 0 & \cdot & \cdot & \cdot & 1 & 0 \\ -c_0/c_k & -c_1/c_k & \cdot & \cdot & \cdot & -c_{k-1}/c_k & 1/c_k \end{bmatrix}$$

Use this expression for $\mathbf{G}^{-1}$ and part (iii) to conclude that $\hat{\theta}$ and its standard error are obtained as the coefficient on x_{tk}/c_k in the regression of

$$y_t \text{ on } [1 - (c_0/c_k)x_{tk}], [x_{t1} - (c_1/c_k)x_{tk}], \ldots, [x_{t,k-1} - (c_{k-1}/c_k)x_{tk}], x_{tk}/c_k, t = 1, \ldots, n.$$

This regression is exactly the one obtained by writing β_k in terms of θ and $\beta_0, \beta_1, \ldots, \beta_{k-1}$, plugging the result into the original model, and rearranging. Therefore, we can formally justify the trick we use throughout the text for obtaining the standard error of a linear combination of parameters.

5 Assume that the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfies the Gauss-Markov assumptions and let $\hat{\boldsymbol{\beta}}$ be the OLS estimator of $\boldsymbol{\beta}$. Let $\mathbf{Z} = \mathbf{G}(\mathbf{X})$ be an $n \times (k+1)$ matrix function of $\mathbf{X}$ and assume that $\mathbf{Z}'\mathbf{X}$[a $(k+1) \times (k+1)$ matrix] is nonsingular. Define a new estimator of $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$.

(i) Show that $\mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$, so that $\tilde{\boldsymbol{\beta}}$ is also unbiased conditional on $\mathbf{X}$.

(ii) Find $\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X})$. Make sure this is a symmetric, $(k+1) \times (k+1)$ matrix that depends on $\mathbf{Z}$, $\mathbf{X}$, and σ^2.

(iii) Which estimator do you prefer, $\hat{\boldsymbol{\beta}}$ or $\tilde{\boldsymbol{\beta}}$? Explain.

6 Consider the setup of the Frisch-Waugh Theorem.

(i) Using partitioned matrices, show that the first order conditions $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ can be written as

$$\mathbf{X}_1'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{y}$$
$$\mathbf{X}_2'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{y}.$$

(ii) Multiply the first set of equations by $\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ and subtract the result from the second set of equations to show that

$$(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{M}_1\mathbf{y},$$

where $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Conclude that

$$\hat{\boldsymbol{\beta}}_2 = (\ddot{\mathbf{X}}_2'\ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}_2'\mathbf{y}.$$

(iii) Use part (ii) to show that

$$\hat{\boldsymbol{\beta}}_2 = (\ddot{\mathbf{X}}_2'\ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}_2'\ddot{\mathbf{y}}.$$

(iv) Use the fact that $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$ to show that the residuals $\ddot{\mathbf{u}}$ from the regression $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{X}}_2$ are identical to the residuals $\hat{\mathbf{u}}$ from the regression $\mathbf{y}$ on $\mathbf{X}_1$, $\mathbf{X}_2$. [*Hint*: By definition and the FW theorem,

$$\ddot{\mathbf{u}} = \ddot{\mathbf{y}} - \ddot{\mathbf{X}}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2) = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2).$$

Now you do the rest.]

7 Suppose that the linear model, written in matrix notation,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

satisfies Assumptions E.1, E.2, and E.3. Partition the model as

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u},$$

where $\mathbf{X}_1$ is $n \times (k_1 + 1)$ and $\mathbf{X}_2$ is $n \times k_2$.

(i) Consider the following proposal for estimating $\boldsymbol{\beta}_2$. First, regress $\mathbf{y}$ on $\mathbf{X}_1$ and obtain the residuals, say, $\ddot{\mathbf{y}}$. Then, regress $\ddot{\mathbf{y}}$ on $\mathbf{X}_2$ to get $\breve{\boldsymbol{\beta}}_2$. Show that $\breve{\boldsymbol{\beta}}_2$ is generally biased and show what the bias is. [You should find $\mathrm{E}(\breve{\boldsymbol{\beta}}_2|\mathbf{X})$ in terms of $\boldsymbol{\beta}_2$, $\mathbf{X}_2$, and the residual-making matrix $\mathbf{M}_1$.]

(ii) As a special case, write

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \beta_k\mathbf{X}_k + \mathbf{u},$$

where $\mathbf{X}_k$ is an $n \times 1$ vector on the variable x_{tk}. Show that

$$\mathrm{E}(\tilde{\beta}_k|\mathbf{X}) = \left(\frac{\mathrm{SSR}_k}{\sum_{t=1}^{n} x_{tk}^2}\right)\beta_k,$$

SSR_k is the sum of squared residuals from regressing x_{tk} on $1, x_{t1}, x_{t2}, \ldots, x_{t,k-1}$. Why is the factor multiplying β_k never greater than one?

(iii) Suppose you know $\boldsymbol{\beta}_1$. Show that the regression $\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1$ on $\mathbf{X}_2$ produces an unbiased estimator of $\boldsymbol{\beta}_2$ (conditional on $\mathbf{X}$).

8 In the context of multiple regression, define the $n \times n$ matrix

$$\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

(i) Show that $\mathbf{M}$ is symmetric and idempotent.

(ii) Prove that m_{tt}, the diagonals of the matrix $\mathbf{M}$, satisfy $0 \leq m_{tt} \leq 1$ for $t = 1, 2, \ldots, n$.

(iii) Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfies the Gauss-Markov Assumptions. Let $\hat{\mathbf{u}}$ be the vector of OLS residuals. Show that

$$\mathrm{E}(\hat{\mathbf{u}}\hat{\mathbf{u}}'|\mathbf{X}) = \sigma^2\mathbf{M}$$

(iv) Conclude that while the errors $\{u_t: t = 1, 2, \ldots, \mathrm{n}\}$ are homoskedastic and uncorrelated under the Gauss-Markov Assumptions, the OLS residuals are heteroskedastic and correlated.

9 Consider the population model

$$\mathrm{y} = \mathbf{x}\boldsymbol{\beta} + u$$
$$\mathrm{E}(u|\mathbf{x}) = \mathbf{0},$$

where the $1 \times (k + 1)$ vector $\mathbf{x}$ is

$$\mathbf{x} = (1, x_1, x_2, \ldots, x_k).$$

Let $\{(\mathbf{x}_i,\mathrm{y}_i): i = 1, 2, \ldots, \mathrm{n}\}$ be a random sample. Show that Assumptions E.3 and E.5 hold.

Answers to Going Further Questions

Chapter 2

Question 2.1: Equation (2.6) would hold when unobserved factors such as student ability, motivation, age, etc., are not related to attendance. In other words, the average value the unobservables (u) should not depend on the value of attendance (x).

Question 2.2: About \$11.05. To see this, from the average wages measured in 1976 and 2003 dollars, the CPI deflator is computed as 19.06/5.90 = 3.23. Multiplying \$3.42 by 3.23 yields about \$11.05.

Question 2.3: The equation will be $salaryhun = 9{,}631.91 + 185.01\ roe$ as is easily seen by multiplying equation (2.39) by 10.

Question 2.4: The equation will be $\widehat{salaryhun} = 9{,}631.91 + 185.01\ roe$, as is easily seen by multiplying equation (2.39) by 10.

Question 2.5: Equation (2.58) can be written as $\text{Var}(\hat{\beta}_0) = (\sigma^2 n^{-1})(\sum_{i=1}^{n} x_i^2)/(\sum_{i=1}^{n}(x_i - \bar{x})^2)$, where the term multiplying $\sigma^2 n^{-1}$ is greater than or equal to one, but it is equal to one if, and only if, $\bar{x} = 0$. In this case, the variance is as small as it can possibly be: $\text{Var}(\hat{\beta}_0) = \sigma^2/n$.

Chapter 3

Question 3.1: Just a few factors include age and gender distribution, size of the police force (or, more generally, resources devoted to crime fighting), population, and general historical factors. These factors certainly might be correlated with *prbconv* and *avgsen*, which means equation (3.5) would not hold. For example, size of the police force is possibly correlated with both *prbcon* and *avgsen*, as some cities put more effort into crime prevention and law enforcement. We should try to bring as many of these factors into the equation as possible.

Question 3.2: About 3.06. Using the third property of OLS concerning predicted values and residuals, plug the average values of all independent variables into the OLS regression line to obtain the average value of the dependent variable. So $\overline{colGPA} = 1.29 + .453\,\overline{hsGPA} + .0094\,\overline{ACT} = 1.29 + .453(3.4) + .0094(24.2) \approx 3.06$. You can check the average of *colGPA* in GPA1 to verify this to the second decimal place.

Question 3.3: No. The variable *shareA* is not an exact linear function of *expendA* and *expendB*, even though it is an exact *nonlinear* function: $shareA = 100 \cdot [expendA/(expendA + expendB)]$. Therefore, it is legitimate to have *expendA*, *expendB*, and *shareA* as explanatory variables.

Question 3.4: As we discussed in Section 3.4, if we are interested in the effect of x_1 on y, correlation among the other explanatory variables (x_2, x_3, and so on) does not affect $\text{Var}(\hat{\beta}_1)$. These variables are included as controls, and we do not have to worry about collinearity among the control variables. Of course, we are controlling for them primarily because we think they are correlated with attendance, but this is necessary to perform a ceteris paribus analysis.

Chapter 4

Question 4.1: Under these assumptions, the Gauss-Markov assumptions are satisfied: u is independent of the explanatory variables, so $\text{E}(u|x_1, \ldots, x_k) = \text{E}(u)$, and $\text{Var}(u|x_1, \ldots, x_k) = \text{Var}(u)$. Further, it is easily seen that $\text{E}(u) = 0$. Therefore, MLR.4 and MLR.5 hold. The classical linear model assumptions are not satisfied because u is not normally distributed (which is a violation of MLR.6).

Question 4.2: $\text{H}_0: \beta_1 = 0, \text{H}_1: \beta_1 < 0$.

Question 4.3: Because $\hat{\beta}_1 = .56 > 0$ and we are testing against $\text{H}_1: \beta_1 > 0$, the one-sided p-value is one-half of the two-sided p-value, or .043.

Question 4.4: $\text{H}_0: \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. $k = 8$ and $q = 4$. The restricted version of the model is

$$score = \beta_0 + \beta_1 classize + \beta_2 expend + \beta_3 tchcomp + \beta_4 enroll + u.$$

Question 4.5: The F statistic for testing exclusion of ACT is $[(.291 - .183)/(1 - .291)](680 - 3) \approx 103.13$. Therefore, the absolute value of the t statistic is about 10.16. The t statistic on ACT is negative, because $\hat{\beta}_{ACT}$ is negative, so $t_{ACT} = -10.16$.

Question 4.6: Not by much. The F test for joint significance of *droprate* and *gradrate* is easily computed from the R-squareds in the table: $F = [(.361 - .353)/(1 - .361)](402/2) \approx 2.52$. The 10% critical value is obtained from Table G.3a as 2.30, while the 5% critical value from Table G.3b is 3. The p-value is about .082. Thus, *droprate* and *gradrate* are jointly significant at the 10% level, but not at the 5% level. In any case, controlling for these variables has a minor effect on the *b*/*s* coefficient.

Chapter 5

Question 5.1: Assuming that $\beta_2 > 0$ (*score* depends positively on *priGPA*) and $\text{Cov}(skipped, priGPA) < 0$ (*skipped* and *priGPA* are negatively correlated), it follows that $\beta_2 \delta_1 < 0$, which means that plim. Because β_1 is thought to be negative (or at least nonpositive), a simple regression is likely to overestimate the importance of skipping classes.

Question 5.2: $\hat{\beta}_j \pm 1.96\text{se}(\hat{\beta}_j)$ is the asymptotic 95% confidence interval. Or, we can replace 1.96 with 2.

Chapter 6

Question 6.1: Because $fincdol = 1{,}000 \cdot faminc$, the coefficient on *fincdol* will be the coefficient on *faminc* divided by 1,000, or .0927/1,000 = .0000927. The standard error also drops by a factor of 1,000, so the t statistic does not change, nor do any of the other OLS statistics. For readability, it is better to measure family income in thousands of dollars.

Question 6.2: Use a more general form of the equation

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \cdots,$$

where x_2 is a proportion rather than a percentage. Then, ceteris paribus, $\Delta\log(y) = \beta_2 \Delta x_2$, $100 \cdot \Delta\log(y) = \beta_2(100 \cdot \Delta x_2)$, or $\%\Delta y \approx \beta_2(100 \cdot \Delta x_2)$. Now, because Δx_2 is the change in the proportion, $100 \cdot \Delta x_2$ is a percentage point change. In particular, if $\Delta x_2 = .01$, then $100 \cdot \Delta x_2 = 1$, which corresponds to a one percentage point change. But then β_2 is the percentage change in y when $100 \cdot \Delta x_2 = 1$.

Question 6.3: The new model would be $stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + \beta_7 ACT \cdot atndrte + u$. Therefore, the partial effect of *atndrte* on *stndfnl* is $\beta_1 + \beta_6 priGPA + \beta_7 \text{ACT}$. This is what we multiply by $\Delta atndrte$ to obtain the ceteris paribus change in *stndfnl*.

Question 6.4: From equation (6.21), $\bar{R}^2 = 1 - \hat{\sigma}^2/[\text{SST}/(n-1)]$. For a given sample and a given dependent variable, $\text{SST}/(n-1)$ is fixed. When we use different sets of explanatory variables, only $\hat{\sigma}^2$ changes. As $\hat{\sigma}^2$ decreases, $\bar{R}^2$ increases. If we make $\hat{\sigma}$, and therefore $\hat{\sigma}^2$, as small as possible, we are making $\bar{R}^2$ as large as possible.

Question 6.5: For a chosen sport—say, players in the National Basketball Association (NBA)—we can collect numerous statistics describing each player's on-court performance. Just a handful of variables include games played, minutes played per game, points scored per game, rebounds per game, assists per game, and measures of defensive efficiency. One has latitude in the actual collection of variables that indicate the productivity of NBA basketball players. Using data on salary and performance, we can run a regression of salary—or, because of the benefits of using the logarithm when a variable is a strictly positive monetary value probably, $lsalary = \log(salary)$—on the measures of performance. The fitted values from the regression give us the predicted log salary based on peformance. Then, we can compute the residuals to see which players have actual *lsalary* above the predicted value (the "overpaid" players) and which have negative residuals (the "underpaid" players). Remember, the residuals always add up to zero. Therefore, by construction, we must find some players are "overpaid" and some are "underpaid."

Chapter 7

Question 7.1: No, because it would not be clear when *party* is one and when it is zero. A better name would be something like *Dem*, which is one for Democratic candidates and zero for Republicans. Or, *Rep*, which is one for Republicans and zero for Democrats.

Question 7.2: With *outfield* as the base group, we would include the dummy variables *frstbase*, *scndbase*, *thrdbase*, *shrtstop*, and *catcher*.

Question 7.3: The null in this case is H_0: $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$, so that there are four restrictions. As usual, we would use an F test (where $q = 4$ and k depends on the number of other explanatory variables).

Question 7.4: Because *tenure* appears as a quadratic, we should allow separate quadratics for men and women. That is, we would add the explanatory variables $female \cdot tenure$ and $female \cdot tenure^2$.

Question 7.5: We plug $pcnv = 0$, $avgsen = 0$, $tottime = 0$, $ptime86 = 0$, $qemp86 = 4$, $black = 1$, and $hispan = 0$ into equation (7.31): $\widehat{arr86} = .380 - .038(4) + .170 = .398$, or almost .4. It is hard to know whether this is "reasonable." For someone with no prior convictions who was employed throughout the year, this estimate might seem high, but remember that the population consists of men who were already arrested at least once prior to 1986.

Chapter 8

Question 8.1: The statement is false. For example, in equation (8.7), the usual standard error for *black* is .147, while the heteroskedasticity-robust standard error is .118.

Question 8.2: The F test would be obtained by regressing $\hat{u}^2$ on *marrmale*, *marrfem*, and *singfem* (*singmale* is the base group). With $n = 526$ and three independent variables in this regression, the *df* are 3 and 522.

Question 8.3: Certainly the outcome of the statistical test suggests some cause for concern. A t statistic of 2.96 is very significant, and it implies that there is heteroskedasticity in the wealth equation. As a practical matter, we know that the WLS standard error, .063, is substantially below the heteroskedasticity-robust standard error for OLS, .104, and so the heteroskedasticity seems to be practically important. (Plus, the nonrobust OLS standard error is .061, which is too optimistic. Therefore, even if we simply adjust the OLS standard error for heteroskedasticity of unknown form, there are nontrivial implications.)

Question 8.4: The 1% critical value in the F distribution with $(2, \infty)$ df is 4.61. An F statistic of 11.15 is well above the 1% critical value, and so we strongly reject the null hypothesis that the transformed errors, $u_i/\sqrt{h_i}$, are homoskedastic. (In fact, the p-value is less than .00002, which is obtained from the $F_{2,804}$ distribution.) This means that our model for $\text{Var}(u|\mathbf{x})$ is inadequate for fully eliminating the heteroskedasticity in u.

Chapter 9

Question 9.1: These are binary variables, and squaring them has no effect: $black^2 = black$, and $hispan^2 = hispan$.

Question 9.2: When *educ*·*IQ* is in the equation, the coefficient on *educ*, say, β_1, measures the effect of *educ* on log(*wage*) when $IQ = 0$. (The partial effect of education is $\beta_1 + \beta_9 IQ$.) There is no one in the population of interest with an IQ close to zero. At the average population IQ, which is 100, the estimated return to education from column (3) is $.018 + .00034(100) = .052$, which is almost what we obtain as the coefficient on *educ* in column (2).

Question 9.3: No. If $educ^*$ is an integer—which means someone has no education past the previous grade completed—the measurement error is zero. If $educ^*$ is not an integer, $educ < educ^*$, so the measurement error is negative. At a minimum, e_1 cannot have zero mean, and e_1 and $educ^*$ are probably correlated.

Question 9.4: An incumbent's decision not to run may be systematically related to how he or she expects to do in the election. Therefore, we may only have a sample of incumbents who are stronger, on average, than all possible incumbents who could run. This results in a sample selection problem if the population of interest includes all incumbents. If we are only interested in the effects of campaign expenditures on election outcomes for incumbents who seek reelection, there is no sample selection problem.

Chapter 10

Question 10.1: The impact propensity is .48, while the long-run propensity is $.48 - .15 + .32 = .65$.

Question 10.2: The explanatory variables are $x_{t1} = z_t$ and $x_{t2} = z_{t-1}$. The absence of perfect collinearity means that these cannot be constant, and there cannot be an exact linear relationship

between them in the sample. This rules out the possibility that all the $z_1, \ldots, z_n$ take on the same value or that the $z_0, z_1, \ldots, z_{n-1}$ take on the same value. But it eliminates other patterns as well. For example, if $z_t = a + bt$ for constants a and b, then $z_{t-1} = a + b(t-1) = (a + bt) - b = z_t - b$, which is a perfect linear function of z_t.

Question 10.3: If $\{z_t\}$ is slowly moving over time—as is the case for the levels or logs of many economic time series—then z_t and z_{t-1} can be highly correlated. For example, the correlation between $unem_t$ and $unem_{t-1}$ in PHILLIPS is .75.

Question 10.4: No, because a linear time trend with $\alpha_1 < 0$ becomes more and more negative as t gets large. Since *gfr* cannot be negative, a linear time trend with a negative trend coefficient cannot represent *gfr* in all future time periods.

Question 10.5: The intercept for March is $\beta_0 + \delta_2$. Seasonal dummy variables are strictly exogenous because they follow a deterministic pattern. For example, the months do not change based upon whether either the explanatory variables or the dependent variables change.

Chapter 11

Question 11.1: (i) No, because $\mathrm{E}(y_t) = \delta_0 + \delta_1 t$ depends on t. (ii) Yes, because $y_t - \mathrm{E}(y_t) = e_t$ is an i.i.d. sequence.

Question 11.2: We plug $inf_t^e = (1/2)inf_{t-1} + (1/2)inf_{t-2}$ into $inf_t - inf_t^e = \beta_1(unem_t - \mu_0) + e_t$ and rearrange: $inf_t - (1/2)(inf_{t-1} + inf_{t-2}) = \beta_0 + \beta_1 unem_t + e_t$, where $\beta_0 = -\beta_1\mu_0$, as before. Therefore, we would regress y_t on $unem_t$, where $y_t = inf_t - (1/2)(inf_{t-1} + inf_{t-2})$. Note that we lose the first two observations in constructing y_t.

Question 11.3: No, because u_t and u_{t-1} are correlated. In particular, $\mathrm{Cov}(u_t, u_{t-1}) = \mathrm{E}[(e_t + \alpha_1 e_{t-1})(e_{t-1} + \alpha_1 e_{t-2})] = \alpha_1 \mathrm{E}(e_{t-1}^2) = \alpha_1\sigma_e^2 \neq 0$ if $\alpha_1 \neq 0$. If the errors are serially correlated, the model cannot be dynamically complete.

Chapter 12

Question 12.1: We use equation (12.4). Now, only adjacent terms are correlated. In particular, the covariance between $x_t u_t$ and $x_{t+1}u_{t+1}$ is $x_t x_{t+1}\mathrm{Cov}(u_t, u_{t+1}) = x_t x_{t+1}\alpha\sigma_e^2$. Therefore, the formula is

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1) &= \mathrm{SST}_x^{-2}\left(\sum_{t=1}^{n} x_t^2 \mathrm{Var}(u_t) + 2\sum_{t=1}^{n-1} x_t x_{t+1} E(u_t u_{t+1})\right) \\ &= \sigma^2/\mathrm{SST}_x + (2/\mathrm{SST}_x^2)\sum_{t=1}^{n-1} \alpha\sigma_e^2 x_t x_{t+1} \\ &= \sigma^2/\mathrm{SST}_x + \alpha\sigma_e^2(2/\mathrm{SST}_x^2)\sum_{t=1}^{n-1} x_t x_{t+1},\end{aligned}$$

where $\sigma^2 = \mathrm{Var}(u_t) = \sigma_e^2 + \alpha_1^2\sigma_e^2 = \sigma_e^2(1 + \alpha_1^2)$. Unless x_t and x_{t+1} are uncorrelated in the sample, the second term is nonzero whenever $\alpha_1 \neq 0$. Notice that if x_t and x_{t+1} are positively correlated and $\alpha < 0$, the true variance is actually *smaller* than the usual variance. When the equation is in levels (as opposed to being differenced), the typical case is $\alpha > 0$, with positive correlation between x_t and x_{t+1}.

Question 12.2: $\hat{\rho} \pm 1.96\mathrm{se}(\hat{\rho})$, where $\mathrm{se}(\hat{\rho})$ is the standard error reported in the regression. Or, we could use the heteroskedasticity-robust standard error. Showing that this is asymptotically valid is complicated because the OLS residuals depend on $\hat{\beta}_j$, but it can be done.

Question 12.3: The model we have in mind is $u_t = \rho_1 u_{t-1} + \rho_4 u_{t-4} + e_t$, and we want to test H_0: $\rho_1 = 0$, $\rho_4 = 0$ against the alternative that H_0 is false. We would run the regression of $\hat{u}_t$ on $\hat{u}_{t-1}$ and $\hat{u}_{t-4}$ to obtain the usual F statistic for joint significance of the two lags. (We are testing two restrictions.)

Question 12.4: We would to estimate the equation using first differences, as $\hat{\rho} = .92$ is close enough to 1 to raise questions about the levels regression. See Chapter 18 for more discussion.

Question 12.5: Because there is only one explanatory variable, the White test can be computed by regressing $\hat{u}_t^2$ on $return_{t-1}$ and $return_{t-1}^2$ (with an intercept, as always) and compute the F test for joint significance of $return_{t-1}$ and $return_{t-1}^2$. If these are jointly significant at a small enough significance level, we reject the null of homoskedasticity.

Chapter 13

Question 13.1: Yes, assuming that we have controlled for all relevant factors. The coefficient on *black* is 1.076, and, with a standard error of .174, it is not statistically different from 1. The 95% confidence interval is from about .735 to 1.417.

Question 13.2: The coefficient on *highearn* shows that, in the absence of any change in the earnings cap, high earners spend much more time—on the order of 29.2% on average [because $\exp(.256) - 1 \approx .292$]—on workers' compensation.

Question 13.3: $E(v_{i1}) = E(a_i + u_{i1}) = E(a_i) + E(v_{i1}) = 0$. Similarly, $E(v_{i2}) = 0$. Therefore, the covariance between v_{i1} and v_{i2} is $E(v_{i1}v_{i2}) = E[(a_i + u_{i1})(a_i + u_{i2})] = E(a_i^2) + E(a_i u_{i1}) + E(a_i u_{i2}) + E(u_{i1}u_{i2}) = E(a_i^2)$, because all of the covariance terms are zero by assumption. But $E(a_i^2) = Var(a_i)$, because $E(a_i) = 0$. This causes positive serial correlation across time in the errors within each i, which biases the usual OLS standard errors in a pooled OLS regression.

Question 13.4: Because $\Delta admn = admn_{90} - admn_{85}$ is the difference in binary indicators, it can be -1 if, and only if, $admn_{90} = 0$ and $admn_{85} = 1$. In other words, Washington state had an administrative per se law in 1985 but it was repealed by 1990.

Question 13.5: No, just as it does not cause bias and inconsistency in a time series regression with strictly exogenous explanatory variables. There are two reasons it is a concern. First, serial correlation in the errors in any equation generally biases the usual OLS standard errors and test statistics. Second, it means that pooled OLS is not as efficient as estimators that account for the serial correlation (as in Chapter 12).

Chapter 14

Question 14.1: Whether we use first differencing or the within transformation, we will have trouble estimating the coefficient on $kids_{it}$. For example, using the within transformation, if $kids_{it}$ does not vary for family i, then $\ddot{kids}_{it} = kids_{it} - \overline{kids}_i = 0$ for $t = 1,2,3$. As long as some families have variation in $kids_{it}$, then we can compute the fixed effects estimator, but the *kids* coefficient could be very imprecisely estimated. This is a form of multicollinearity in fixed effects estimation (or first-differencing estimation).

Question 14.2: If a firm did not receive a grant in the first year, it may or may not receive a grant in the second year. But if a firm did receive a grant in the first year, it could not get a grant in

the second year. That is, if $grant_{-1} = 1$, then $grant = 0$. This induces a negative correlation between $grant$ and $grant_{-1}$. We can verify this by computing a regression of $grant$ on $grant_{-1}$, using the data in JTRAIN for 1989. Using all firms in the sample, we get

$$\widehat{grant} = .248 - .248\ grant_{-1}$$
$$(.035)\quad (.072)$$
$$n = 157, R^2 = .070.$$

The coefficient on $grant_{-1}$ must be the negative of the intercept because $\widehat{grant} = 0$ when $grant_{-1} = 1$.

Question 14.3: It suggests that the unobserved effect a_i is positively correlated with $union_{it}$. Remember, pooled OLS leaves a_i in the error term, while fixed effects removes a_i. By definition, a_i has a positive effect on log(*wage*). By the standard omitted variables analysis (see Chapter 3), OLS has an upward bias when the explanatory variable (*union*) is positively correlated with the omitted variable (a_i). Thus, belonging to a union appears to be positively related to time-constant, unobserved factors that affect wage.

Question 14.4: Not if all sisters within a family have the same mother and father. Then, because the parents' race variables would not change by sister, they would be differenced away in equation (14.13).

Chapter 15

Question 15.1: Probably not. In the simple equation (15.18), years of education is part of the error term. If some men who were assigned low draft lottery numbers obtained additional schooling, then lottery number and education are negatively correlated, which violates the first requirement for an instrumental variable in equation (15.4).

Question 15.2: (i) For equation (15.27), we require that high school peer group effects carry over to college. Namely, for a given SAT score, a student who went to a high school where smoking marijuana was more popular would smoke more marijuana in college. Even if the identification condition equation (15.27) holds, the link might be weak.

(ii) We have to assume that percentage of students using marijuana at a student's high school is not correlated with unobserved factors that affect college grade point average. Although we are somewhat controlling for high school quality by including *SAT* in the equation, this might not be enough. Perhaps high schools that did a better job of preparing students for college also had fewer students smoking marijuana. Or marijuana usage could be correlated with average income levels. These are, of course, empirical questions that we may or may not be able to answer.

Question 15.3: Although prevalence of the NRA and subscribers to gun magazines are probably correlated with the presence of gun control legislation, it is not obvious that they are uncorrelated with unobserved factors that affect the violent crime rate. In fact, we might argue that a population interested in guns is a reflection of high crime rates, and controlling for economic and demographic variables is not sufficient to capture this. It would be hard to argue persuasively that these are truly exogenous in the violent crime equation.

Question 15.4: As usual, there are two requirements. First, it should be the case that growth in government spending is systematically related to the party of the president, after netting out the investment rate and growth in the labor force. In other words, the instrument must be partially correlated with the endogenous explanatory variable. While we might think that government spending grows more slowly under Republican presidents, this certainly has not always been true in the United States and would have to be tested using the t statistic on REP_{t-1} in the reduced form

$gGov_t = \pi_0 + \pi_1 REP_{t-1} + \pi_2 INVRAT_t + \pi_3 gLAB_t + v_t$. We must assume that the party of the president has no separate effect on *gGDP*. This would be violated if, for example, monetary policy differs systematically by presidential party and has a separate effect on GDP growth.

Chapter 16

Question 16.1: Probably not. It is because firms choose price and advertising expenditures jointly that we are not interested in the experiment where, say, advertising changes exogenously and we want to know the effect on price. Instead, we would model price and advertising each as a function of demand and cost variables. This is what falls out of the economic theory.

Question 16.2: We must assume two things. First, money supply growth should appear in equation (16.22), so that it is partially correlated with *inf*. Second, we must assume that money supply growth does not appear in equation (16.23). If we think we must include money supply growth in equation (16.23), then we are still short an instrument for *inf*. Of course, the assumption that money supply growth is exogenous can also be questioned.

Question 16.3: Use the Hausman test from Chapter 15. In particular, let $\hat{v}_2$ be the OLS residuals from the reduced form regression of *open* on log(*pcinc*) and log(*land*). Then, use an OLS regression of *inf* on *open*, log(*pcinc*), and $\hat{v}_2$ and compute the *t* statistic for significance of $\hat{v}_2$. If $\hat{v}_2$ is significant, the 2SLS and OLS estimates are statistically different.

Question 16.4: The demand equation looks like

$$\log(fish_t) = \beta_0 + \beta_1 \log(prcfish_t) + \beta_2 \log(inc_t) + \beta_3 \log(prcchick_t) + \beta_4 \log(prcbeef_t) + u_{t1},$$

where logarithms are used so that all elasticities are constant. By assumption, the demand function contains no seasonality, so the equation does not contain monthly dummy variables (say, feb_t, mar_t, …, dec_t, with January as the base month). Also, by assumption, the supply of fish is seasonal, which means that the supply function does depend on at least some of the monthly dummy variables. Even without solving the reduced form for log(*prcfish*), we conclude that it depends on the monthly dummy variables. Since these are exogenous, they can be used as instruments for log(*prcfish*) in the demand equation. Therefore, we can estimate the demand-for-fish equation using monthly dummies as the IVs for log(*prcfish*). Identification requires that at least one monthly dummy variable appears with a nonzero coefficient in the reduced form for log(*prcfish*).

Chapter 17

Question 17.1: H_0: $\beta_4 = \beta_5 = \beta_6 = 0$, so that there are three restrictions and therefore three *df* in the *LR* or Wald test.

Question 17.2: We need the partial derivative of $\Phi(\hat{\beta}_0 + \hat{\beta}_1 nwifeinc + \hat{\beta}_2 educ + \hat{\beta}_3 exper + \hat{\beta}_4 exper^2 + \cdots)$ with respect to *exper*, which is $\phi(\cdot)(\hat{\beta}_3 + 2\hat{\beta}_4 exper)$, where $\phi(\cdot)$ is evaluated at the given values and the initial level of experience. Therefore, we need to evaluate the standard normal probability density at $.270 - .012(20.13) + .131(12.3) + .123(10) - .0019(10^2) - .053(42.5) - .868(0) + .036(1) \approx .463$, where we plug in the initial level of experience (10). But $\phi(.463) = (2\pi)^{-1/2}\exp[-(.463^2)/2] \approx .358$. Next, we multiply this by $\hat{\beta}_3 + 2\hat{\beta}_4 exper$, which is evaluated at $exper = 10$. The partial effect using the calculus approximation is $.358[.123 - 2(.0019)(10)] \approx .030$. In other words, at the given values of the explanatory variables and starting at $exper = 10$, the next year of experience increases the probability of labor force participation by about .03.

Question 17.3: No. The number of extramarital affairs is a nonnegative integer, which presumably takes on zero or small numbers for a substantial fraction of the population. It is not realistic to use a Tobit model, which, while allowing a pileup at zero, treats y as being continuously distributed over positive values. Formally, assuming that $y = \max(0, y^*)$, where y^* is normally distributed, is at odds with the discreteness of the number of extramarital affairs when $y > 0$.

Question 17.4: The adjusted standard errors are the usual Poisson MLE standard errors multiplied by $\hat{\sigma} = \sqrt{2} \approx 1.41$, so the adjusted standard errors will be about 41% higher. The quasi-*LR* statistic is the usual *LR* statistic divided by $\hat{\sigma}^2$, so it will be one-half of the usual *LR* statistic.

Question 17.5: By assumption, $mvp_i = \beta_0 + \mathbf{x}_i\boldsymbol{\beta} + u_i$, where, as usual, $\mathbf{x}_i\boldsymbol{\beta}$ denotes a linear function of the exogenous variables. Now, observed wage is the largest of the minimum wage and the marginal value product, so $wage_i = \max(minwage_i, mvp_i)$, which is very similar to equation (17.34), except that the max operator has replaced the min operator.

Chapter 18

Question 18.1: We can plug these values directly into equation (18.1) and take expectations. First, because $z_s = 0$, for all $s < 0$, $y_{-1} = \alpha + u_{-1}$. Then, $z_0 = 1$, so $y_0 = \alpha + \delta_0 + u_0$. For $h \geq 1$, $y_h = \alpha + \delta_{h-1} + \delta_h + u_h$. Because the errors have zero expected values, $\mathrm{E}(y_{-1}) = \alpha$, $\mathrm{E}(y_0) = \alpha + \delta_0$, and $\mathrm{E}(y_h) = \alpha + \delta_{h-1} + \delta$, for all $h \geq 1$. As $h \to \infty$, $\delta_h \to 0$. It follows that $\mathrm{E}(y_h) \to \alpha$ as $h \to \infty$, that is, the expected value of y_h returns to the expected value before the increase in z, at time zero. This makes sense: although the increase in z lasted for two periods, it is still a temporary increase.

Question 18.2: Under the described setup, Δy_t and Δx_t are i.i.d. sequences that are independent of one another. In particular, Δy_t and Δx_t are uncorrelated. If $\hat{\gamma}_t$ is the slope coefficient from regressing Δy_t on Δx_t, $t = 1, 2, \ldots, n$, then plim $\hat{\gamma}_t = 0$. This is as it should be, as we are regressing one I(0) process on another I(0) process, and they are uncorrelated. We write the equation $\Delta y_t = \gamma_0 + \gamma_1 \Delta x_t + e_t$, where $\gamma_0 = \gamma_1 = 0$. Because $\{e_t\}$ is independent of $\{\Delta x_t\}$, the strict exogeneity assumption holds. Moreover, $\{e_t\}$ is serially uncorrelated and homoskedastic. By Theorem 11.2 in Chapter 11, the t statistic for $\hat{\gamma}_t$ has an approximate standard normal distribution. If e_t is normally distributed, the classical linear model assumptions hold, and the t statistic has an exact t distribution.

Question 18.3: Write $x_t = x_{t-1} + a_t$, where $\{a_t\}$ is I(0). By assumption, there is a linear combination, say, $s_t = y_t - \beta x_t$, which is I(0). Now, $y_t - \beta x_{t-1} = y_t - \beta(x_t - \alpha_t) = s_t + \beta a_t$. Because s_t and a_t are I(0) by assumption, so is $s_t + \beta a_t$.

Question 18.4: Just use the sum of squared residuals form of the F test and assume homoskedasticity. The restricted SSR is obtained by regressing $\Delta hy6_t - \Delta hy3_{t-1} + (hy6_{t-1} - hy3_{t-2})$ on a constant. Notice that a_0 is the only parameter to estimate in $\Delta hy6_t = \alpha_0 + \gamma_0 \Delta hy3_{t-1} + \delta(hy6_{t-1} - hy3_{t-2})$ when the restrictions are imposed. The unrestricted sum of squared residuals is obtained from equation (18.39).

Question 18.5: We are fitting two equations: $\hat{y}_t = \hat{\alpha} + \hat{\beta}t$ and $\hat{y}_t = \hat{\gamma} + \hat{\delta} year_t$. We can obtain the relationship between the parameters by noting that $year_t = t + 49$. Plugging this into the second equation gives $\hat{y}_t = \hat{\gamma} + \hat{\delta}(t + 49) = (\hat{\gamma} + 49\hat{\delta}) + \hat{\delta}t$. Matching the slope and intercept with the first equation gives $\hat{\delta} = \hat{\beta}$—so that the slopes on t and $year_t$ are identical—and $\hat{\alpha} = \hat{\gamma} + 49\hat{\delta}$. Generally, when we use *year* rather than t, the intercept will change, but the slope will not. (You can verify this by using one of the time series data sets, such as HSEINV or INVEN.) Whether we use t or some measure of *year* does not change fitted values, and, naturally, it does not change forecasts of future values. The intercept simply adjusts appropriately to different ways of including a trend in the regression.

Statistical Tables

TABLE G.1 Cumulative Areas under the Standard Normal Distribution

z	0	1	2	3	4	5	6	7	8	9
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121

(*continued*)

TABLE G.1 (*Continued*)

z	0	1	2	3	4	5	6	7	8	9
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Examples: If $Z \sim \text{Normal}(0, 1)$, then $\text{P}(Z \leq -1.32) = .0934$ and $\text{P}(Z \leq 1.84) = .9671$.

Source: This table was generated using the Stata® function normal.

TABLE G.2 Critical Values of the *t* Distribution

		Significance Level				
1-Tailed:		.10	.05	.025	.01	.005
2-Tailed:		.20	.10	.05	.02	.01
Degrees of Freedom	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	40	1.303	1.684	2.021	2.423	2.704
	60	1.296	1.671	2.000	2.390	2.660
	90	1.291	1.662	1.987	2.368	2.632
	120	1.289	1.658	1.980	2.358	2.617
	∞	1.282	1.645	1.960	2.326	2.576

Examples: The 1% critical value for a one-tailed test with 25 *df* is 2.485. The 5% critical value for a two-tailed test with large (> 120) *df* is 1.96.

Source: This table was generated using the Stata® function invttail.

TABLE G.3a 10% Critical Values of the *F* Distribution

Denominator Degrees of Freedom	Numerator Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60

Example: The 10% critical value for numerator $df = 2$ and denominator $df = 40$ is 2.44.
Source: This table was generated using the Stata® function invFtail.

TABLE G.3b 5% Critical Values of the *F* Distribution

		Numerator Degrees of Freedom									
		1	2	3	4	5	6	7	8	9	10
Denominator Degrees of Freedom	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Example: The 5% critical value for numerator $df = 4$ and large denominator $df(\infty)$ is 2.37.

Source: This table was generated using the Stata® function invFtail.

TABLE G.3c 1% Critical Values of the *F* Distribution

		Numerator Degrees of Freedom									
		1	2	3	4	5	6	7	8	9	10
Denominator Degrees of Freedom	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Example: The 1% critical value for numerator $df = 3$ and denominator $df = 60$ is 4.13.

Source: This table was generated using the Stata® function invFtail.

TABLE G.4 Critical Values of the Chi-Square Distribution

		Significance Level		
		.10	.05	.01
Degrees of Freedom	1	2.71	3.84	6.63
	2	4.61	5.99	9.21
	3	6.25	7.81	11.34
	4	7.78	9.49	13.28
	5	9.24	11.07	15.09
	6	10.64	12.59	16.81
	7	12.02	14.07	18.48
	8	13.36	15.51	20.09
	9	14.68	16.92	21.67
	10	15.99	18.31	23.21
	11	17.28	19.68	24.72
	12	18.55	21.03	26.22
	13	19.81	22.36	27.69
	14	21.06	23.68	29.14
	15	22.31	25.00	30.58
	16	23.54	26.30	32.00
	17	24.77	27.59	33.41
	18	25.99	28.87	34.81
	19	27.20	30.14	36.19
	20	28.41	31.41	37.57
	21	29.62	32.67	38.93
	22	30.81	33.92	40.29
	23	32.01	35.17	41.64
	24	33.20	36.42	42.98
	25	34.38	37.65	44.31
	26	35.56	38.89	45.64
	27	36.74	40.11	46.96
	28	37.92	41.34	48.28
	29	39.09	42.56	49.59
	30	40.26	43.77	50.89

Example: The 5% critical value with $df = 8$ is 15.51.

Source: This table was generated using the Stata® function invchi2tail.

References

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review* 80, 313–336.

Angrist, J. D., and A. B. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106, 979–1014.

Ashenfelter, O., and A. B. Krueger (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review* 84, 1157–1173.

Averett, S., and S. Korenman (1996), "The Economic Reality of the Beauty Myth," *Journal of Human Resources* 31, 304–330.

Ayres, I., and S. D. Levitt (1998), "Measuring Positive Externalities from Unobservable Victim Precaution: An Empirical Analysis of Lojack," *Quarterly Journal of Economics* 108, 43–77.

Banerjee, A., J. Dolado, J. W. Galbraith, and D. F. Hendry (1993), *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

Bartik, T. J. (1991), "The Effects of Property Taxes and Other Local Public Policies on the Intrametropolitan Pattern of Business Location," in *Industry Location and Public Policy*, ed. H. W. Herzog and A. M. Schlottmann, 57–80. Knoxville: University of Tennessee Press.

Becker, G. S. (1968), "Crime and Punishment: An Economic Approach," *Journal of Political Economy* 76, 169–217.

Belsley, D., E. Kuh, and R. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Berk, R. A. (1990), "A Primer on Robust Regression," in *Modern Methods of Data Analysis*, ed. J. Fox and J. S. Long, 292–324. Newbury Park, CA: Sage Publications.

Betts, J. R. (1995), "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth," *Review of Economics and Statistics* 77, 231–250.

Biddle, J. E., and D. S. Hamermesh (1990), "Sleep and the Allocation of Time," *Journal of Political Economy* 98, 922–943.

Biddle, J. E., and D. S. Hamermesh (1998), "Beauty, Productivity, and Discrimination: Lawyers' Looks and Lucre," *Journal of Labor Economics* 16, 172–201.

Blackburn, M., and D. Neumark (1992), "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials," *Quarterly Journal of Economics* 107, 1421–1436.

Blinder, A. S., and M. W. Watson (2014), "Presidents and the U.S. Economy: An Econometric Exploration," National Bureau of Economic Research Working Paper No. 20324.

Blomström, M., R. E. Lipsey, and M. Zejan (1996), "Is Fixed Investment the Key to Economic Growth?" *Quarterly Journal of Economics* 111, 269–276.

Blundell, R., A. Duncan, and K. Pendakur (1998), "Semiparametric Estimation and Consumer Demand," *Journal of Applied Econometrics* 13, 435–461.

Bollerslev, T., R. Y. Chou, and K. F. Kroner (1992), "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence," *Journal of Econometrics* 52, 5–59.

Bollerslev, T., R. F. Engle, and D. B. Nelson (1994), "ARCH Models," in *Handbook of Econometrics*, volume 4, chapter 49, ed. R. F. Engle and D. L. McFadden, 2959–3038. Amsterdam: North-Holland.

Bound, J., D. A. Jaeger, and R. M. Baker (1995), "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and Endogenous Explanatory Variables Is Weak," *Journal of the American Statistical Association* 90, 443–450.

Breusch, T. S., and A. R. Pagan (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation," *Econometrica* 47, 987–1007.

Cameron, A. C., and P. K. Trivedi (1998), *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Campbell, J. Y., and N. G. Mankiw (1990), "Permanent Income, Current Income, and Consumption," *Journal of Business and Economic Statistics* 8, 265–279.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, ed. L. N. Christophides, E. K. Grant, and R. Swidinsky, 201–222. Toronto: University of Toronto Press.

Card, D., and A. Krueger (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy* 100, 1–40.

Castillo-Freeman, A. J., and R. B. Freeman (1992), "When the Minimum Wage Really Bites: The Effect of the U.S.-Level Minimum on Puerto Rico," in *Immigration and the Work Force*, ed. G. J. Borjas and R. B. Freeman, 177–211. Chicago: University of Chicago Press.

Clark, K. B. (1984), "Unionization and Firm Performance: The Impact on Profits, Growth, and Productivity," *American Economic Review* 74, 893–919.

Cloninger, D. O. (1991), "Lethal Police Response as a Crime Deterrent: 57-City Study Suggests a Decrease in Certain Crimes," *American Journal of Economics and Sociology* 50, 59–69.

Cloninger, D. O., and L. C. Sartorius (1979), "Crime Rates, Clearance Rates and Enforcement Effort: The Case of Houston, Texas," *American Journal of Economics and Sociology* 38, 389–402.

Cochrane, J. H. (1997), "Where Is the Market Going? Uncertain Facts and Novel Theories," *Economic Perspectives* 21, Federal Reserve Bank of Chicago, 3–37.

Cornwell, C., and W. N. Trumbull (1994), "Estimating the Economic Model of Crime Using Panel Data," *Review of Economics and Statistics* 76, 360–366.

Craig, B. R., W. E. Jackson III, and J. B. Thomson (2007), "Small Firm Finance, Credit Rationing, and the Impact of SBA-Guaranteed Lending on Local Economic Growth," *Journal of Small Business Management* 45, 116–132.

Currie, J. (1995), *Welfare and the Well-Being of Children*. Chur, Switzerland: Harwood Academic Publishers.

Currie, J., and N. Cole (1993), "Welfare and Child Health: The Link between AFDC Participation and Birth Weight," *American Economic Review* 83, 971–983.

Currie, J., and D. Thomas (1995), "Does Head Start Make a Difference?" *American Economic Review* 85, 341–364.

Davidson, R., and J. G. MacKinnon (1981), "Several Tests of Model Specification in the Presence of Alternative Hypotheses," *Econometrica* 49, 781–793.

Davidson, R., and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics*. New York: Oxford University Press.

De Long, J. B., and L. H. Summers (1991), "Equipment Investment and Economic Growth," *Quarterly Journal of Economics* 106, 445–502.

Dickey, D. A., and W. A. Fuller (1979), "Distributions of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association* 74, 427–431.

Diebold, F. X. (2001), *Elements of Forecasting*. 2nd ed. Cincinnati: South-Western.

Downes, T. A., and S. M. Greenstein (1996), "Understanding the Supply Decisions of Nonprofits: Modeling the Location of Private Schools," *Rand Journal of Economics* 27, 365–390.

Draper, N., and H. Smith (1981), *Applied Regression Analysis*. 2nd ed. New York: Wiley.

Duan, N. (1983), "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association* 78, 605–610.

Durbin, J. (1970), "Testing for Serial Correlation in Least Squares Regressions When Some of the Regressors Are Lagged Dependent Variables," *Econometrica* 38, 410–421.

Durbin, J., and G. S. Watson (1950), "Testing for Serial Correlation in Least Squares Regressions I," *Biometrika* 37, 409–428.

Eicker, F. (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 59–82. Berkeley: University of California Press.

Eide, E. (1994), *Economics of Crime: Deterrence and the Rational Offender*. Amsterdam: North-Holland.

Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica* 50, 987–1007.

Engle, R. F., and C. W. J. Granger (1987), "Cointegration and Error Correction: Representation, Estimation, and Testing," *Econometrica* 55, 251–276.

Evans, W. N., and R. M. Schwab (1995), "Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *Quarterly Journal of Economics* 110, 941–974.

Fair, R. C. (1996), "Econometrics and Presidential Elections," *Journal of Economic Perspectives* 10, 89–102.

Franses, P. H., and R. Paap (2001), *Quantitative Models in Marketing Research*. Cambridge: Cambridge University Press.

Freeman, D. G. (2007), "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws," *Contemporary Economic Policy* 25, 293–308.

Friedman, B. M., and K. N. Kuttner (1992), "Money, Income, Prices, and Interest Rates," *American Economic Review* 82, 472–492.

Geronimus, A. T., and S. Korenman (1992), "The Socioeconomic Consequences of Teen Childbearing Reconsidered," *Quarterly Journal of Economics* 107, 1187–1214.

Goldberger, A. S. (1991), *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Graddy, K. (1995), "Testing for Imperfect Competition at the Fulton Fish Market," *Rand Journal of Economics* 26, 75–92.

Graddy, K. (1997), "Do Fast-Food Chains Price Discriminate on the Race and Income Characteristics of an Area?" *Journal of Business and Economic Statistics* 15, 391–401.

Granger, C. W. J., and P. Newbold (1974), "Spurious Regressions in Econometrics," *Journal of Econometrics* 2, 111–120.

Greene, W. (1997), *Econometric Analysis*. 3rd ed. New York: MacMillan.

Griliches, Z. (1957), "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics* 39, 8–20.

Grogger, J. (1990), "The Deterrent Effect of Capital Punishment: An Analysis of Daily Homicide Counts," *Journal of the American Statistical Association* 410, 295–303.

Grogger, J. (1991), "Certainty vs. Severity of Punishment," *Economic Inquiry* 29, 297–309.

Hall, R. E. (1988), "The Relation between Price and Marginal Cost in U.S. Industry," *Journal of Political Economy* 96, 921–948.

Hamermesh, D. S., and J. E. Biddle (1994), "Beauty and the Labor Market," *American Economic Review* 84, 1174–1194.

Hamermesh, D. H., and A. Parker (2005), "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity," *Economics of Education Review* 24, 369–376.

Hamilton, J. D. (1994), *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, C.B. (2007), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T Is Large," *Journal of Econometrics* 141, 597–620.

Hanushek, E. (1986), "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature* 24, 1141–1177.

Harvey, A. (1990), *The Econometric Analysis of Economic Time Series*. 2nd ed. Cambridge, MA: MIT Press.

Hausman, J. A. (1978), "Specification Tests in Econometrics," *Econometrica* 46, 1251–1271.

Hausman, J. A., and D. A. Wise (1977), "Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica* 45, 319–339.

Hayasyi, F. (2000), *Econometrics*. Princeton, NJ: Princeton University Press.

Heckman, J. J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5, 475–492.

Herrnstein, R. J., and C. Murray (1994), *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Hersch, J., and L. S. Stratton (1997), "Housework, Fixed Effects, and Wages of Married Workers," *Journal of Human Resources* 32, 285–307.

Hines, J. R. (1996), "Altered States: Taxes and the Location of Foreign Direct Investment in America," *American Economic Review* 86, 1076–1094.

Holzer, H. (1991), "The Spatial Mismatch Hypothesis: What Has the Evidence Shown?" *Urban Studies* 28, 105–122.

Holzer, H., R. Block, M. Cheatham, and J. Knott (1993), "Are Training Subsidies Effective? The Michigan Experience," *Industrial and Labor Relations Review* 46, 625–636.

Horowitz, J. (2001), "The Bootstrap," in *Handbook of Econometrics*, volume 5, chapter 52, ed. E. Leamer and J. L. Heckman, 3159–3228. Amsterdam: North Holland.

Hoxby, C. M. (1994), "Do Private Schools Provide Competition for Public Schools?" National Bureau of Economic Research Working Paper Number 4978.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 221–233. Berkeley: University of California Press.

Hunter, W. C., and M. B. Walker (1996), "The Cultural Affinity Hypothesis and Mortgage Lending Decisions," *Journal of Real Estate Finance and Economics* 13, 57–70.

Hylleberg, S. (1992), *Modelling Seasonality*. Oxford: Oxford University Press.

Kane, T. J., and C. E. Rouse (1995), "Labor-Market Returns to Two- and Four-Year Colleges," *American Economic Review* 85, 600–614.

Kiefer, N. M., and T. J. Vogelsang (2005), "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory* 21, 1130–1164.

Kiel, K. A., and K. T. McClain (1995), "House Prices during Siting Decision Stages: The Case of an Incinerator from Rumor through Operation," *Journal of Environmental Economics and Management* 28, 241–255.

Kleck, G., and E. B. Patterson (1993), "The Impact of Gun Control and Gun Ownership Levels on Violence Rates," *Journal of Quantitative Criminology* 9, 249–287.

Koenker, R. (1981), "A Note on Studentizing a Test for Heteroskedasticity," *Journal of Econometrics* 17, 107–112.

Koenker, R. (2005), *Quantile Regression*. Cambridge: Cambridge University Press.

Korenman, S., and D. Neumark (1991), "Does Marriage Really Make Men More Productive?" *Journal of Human Resources* 26, 282–307.

Korenman, S., and D. Neumark (1992), "Marriage, Motherhood, and Wages," *Journal of Human Resources* 27, 233–255.

Krueger, A. B. (1993), "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984–1989," *Quarterly Journal of Economics* 108, 33–60.

Krupp, C. M., and P. S. Pollard (1996), "Market Responses to Antidumping Laws: Some Evidence from the U.S. Chemical Industry," *Canadian Journal of Economics* 29, 199–227.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin (1992), "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?" *Journal of Econometrics* 54, 159–178.

Lalonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76, 604–620.

Larsen, R. J., and M. L. Marx (1986), *An Introduction to Mathematical Statistics and Its Applications*. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.

Leamer, E. E. (1983), "Let's Take the Con Out of Econometrics," *American Economic Review* 73, 31–43.

Levine, P. B., A. B. Trainor, and D. J. Zimmerman (1996), "The Effect of Medicaid Abortion Funding Restrictions on Abortions, Pregnancies, and Births," *Journal of Health Economics* 15, 555–578.

Levine, P. B., and D. J. Zimmerman (1995), "The Benefit of Additional High-School Math and Science Classes for Young Men and Women," *Journal of Business and Economics Statistics* 13, 137–149.

Levitt, S. D. (1994), "Using Repeat Challengers to Estimate the Effect of Campaign Spending on Election Outcomes in the U.S. House," *Journal of Political Economy* 102, 777–798.

Levitt, S. D. (1996), "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Legislation," *Quarterly Journal of Economics* 111, 319–351.

Little, R. J. A., and D. B. Rubin (2002), *Statistical Analysis with Missing Data*. 2nd ed. Wiley: New York.

Low, S. A., and L. R. McPheters (1983), "Wage Differentials and the Risk of Death: An Empirical Analysis," *Economic Inquiry* 21, 271–280.

Lynch, L. M. (1992), "Private Sector Training and the Earnings of Young Workers," *American Economic Review* 82, 299–312.

MacKinnon, J. G., and H. White (1985), "Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics* 29, 305–325.

Maloney, M. T., and R. E. McCormick (1993), "An Examination of the Role that Intercollegiate Athletic Participation Plays in Academic Achievement: Athletes' Feats in the Classroom," *Journal of Human Resources* 28, 555–570.

Mankiw, N. G. (1994), *Macroeconomics*. 2nd ed. New York: Worth.

Mark, S. T., T. J. McGuire, and L. E. Papke (2000), "The Influence of Taxes on Employment and Population Growth: Evidence from the Washington, D.C. Metropolitan Area," *National Tax Journal* 53, 105–123.

McCarthy, P. S. (1994), "Relaxed Speed Limits and Highway Safety: New Evidence from California," *Economics Letters* 46, 173–179.

McClain, K. T., and J. M. Wooldridge (1995), "A Simple Test for the Consistency of Dynamic Linear Regression in Rational Distributed Lag Models," *Economics Letters* 48, 235–240.

McCormick, R. E., and M. Tinsley (1987), "Athletics versus Academics: Evidence from SAT Scores," *Journal of Political Economy* 95, 1103–1116.

McFadden, D. L. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. P. Zarembka, 105–142. New York: Academic Press.

Meyer, B. D. (1995), "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics* 13, 151–161.

Meyer, B. D., W. K. Viscusi, and D. L. Durbin (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review* 85, 322–340.

Mizon, G. E., and J. F. Richard (1986), "The Encompassing Principle and Its Application to Testing Nonnested Hypotheses," *Econometrica* 54, 657–678.

Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica* 55, 765–799.

Mullahy, J., and P. R. Portney (1990), "Air Pollution, Cigarette Smoking, and the Production of Respiratory Health," *Journal of Health Economics* 9, 193–205.

Mullahy, J., and J. L. Sindelar (1994), "Do Drinkers Know When to Say When? An Empirical Analysis of Drunk Driving," *Economic Inquiry* 32, 383–394.

Netzer, D. (1992), "Differences in Reliance on User Charges by American State and Local Governments," *Public Finance Quarterly* 20, 499–511.

Neumark, D. (1996), "Sex Discrimination in Restaurant Hiring: An Audit Study," *Quarterly Journal of Economics* 111, 915–941.

Neumark, D., and W. Wascher (1995), "Minimum Wage Effects on Employment and School Enrollment," *Journal of Business and Economic Statistics* 13, 199–206.

Newey, W. K., and K. D. West (1987), "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, 703–708.

Papke, L. E. (1987), "Subnational Taxation and Capital Mobility: Estimates of Tax-Price Elasticities," *National Tax Journal* 40, 191–203.

Papke, L. E. (1994), "Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program," *Journal of Public Economics* 54, 37–49.

Papke, L. E. (1995), "Participation in and Contributions to 401(k) Pension Plans: Evidence from Plan Data," *Journal of Human Resources* 30, 311–325.

Papke, L. E. (1999), "Are 401(k) Plans Replacing Other Employer-Provided Pensions? Evidence from Panel Data," *Journal of Human Resources*, 34, 346–368.

Papke, L. E. (2005), "The Effects of Spending on Test Pass Rates: Evidence from Michigan," *Journal of Public Economics* 89, 821–839.

Papke, L. E., and J. M. Wooldridge (1996), "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates," *Journal of Applied Econometrics* 11, 619–632.

Park, R. (1966), "Estimation with Heteroskedastic Error Terms," *Econometrica* 34, 888.

Peek, J. (1982), "Interest Rates, Income Taxes, and Anticipated Inflation," *American Economic Review* 72, 980–991.

Pindyck, R. S., and D. L. Rubinfeld (1992), *Microeconomics*. 2nd ed. New York: Macmillan.

Ram, R. (1986), "Government Size and Economic Growth: A New Framework and Some Evidence from Cross-Section and Time-Series Data," *American Economic Review* 76, 191–203.

Ramanathan, R. (1995), *Introductory Econometrics with Applications*. 3rd ed. Fort Worth: Dryden Press.

Ramey, V. (1991), "Nonconvex Costs and the Behavior of Inventories," *Journal of Political Economy* 99, 306–334.

Ramsey, J. B. (1969), "Tests for Specification Errors in Classical Linear Least-Squares Analysis," *Journal of the Royal Statistical Association*, Series B, 71, 350–371.

Romer, D. (1993), "Openness and Inflation: Theory and Evidence," *Quarterly Journal of Economics* 108, 869–903.

Rose, N. L. (1985), "The Incidence of Regulatory Rents in the Motor Carrier Industry," *Rand Journal of Economics* 16, 299–318.

Rose, N. L., and A. Shepard (1997), "Firm Diversification and CEO Compensation: Managerial Ability or Executive Entrenchment?" *Rand Journal of Economics* 28, 489–514.

Rouse, C. E. (1998), "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *Quarterly Journal of Economics* 113, 553–602.

Sander, W. (1992), "The Effect of Women's Schooling on Fertility," *Economic Letters* 40, 229–233.

Savin, N. E., and K. J. White (1977), "The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors," *Econometrica* 45, 1989–1996.

Shea, J. (1993), "The Input-Output Approach to Instrument Selection," *Journal of Business and Economic Statistics* 11, 145–155.

Shughart, W. F., and R. D. Tollison (1984), "The Random Character of Merger Activity," *Rand Journal of Economics* 15, 500–509.

Solon, G. (1985), "The Minimum Wage and Teenage Employment: A Re-analysis with Attention to Serial Correlation and Seasonality," *Journal of Human Resources* 20, 292–297.

Staiger, D., and J. H. Stock (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica* 65, 557–586.

Stigler, S. M. (1986), *The History of Statistics*. Cambridge, MA: Harvard University Press.

Stock, J. H., and M. W. Watson (1989), "Interpreting the Evidence on Money-Income Causality," *Journal of Econometrics* 40, 161–181.

Stock, J. H., and M. W. Watson (1993), "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica* 61, 783–820.

Stock, J. H., and M. Yogo (2005), "Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. D. W. K. Andrews and J. H. Stock, 109–120. Cambridge: Cambridge University Press.

Stock, J. W., and M. W. Watson (2008), "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression," *Econometrica* 76, 155–174.

Sydsaeter, K., and P. J. Hammond (1995), *Mathematics for Economic Analysis*. Englewood Cliffs, NJ: Prentice Hall.

Terza, J. V. (2002), "Alcohol Abuse and Employment: A Second Look," *Journal of Applied Econometrics* 17, 393–404.

Tucker, I. B. (2004), "A Reexamination of the Effect of Big-time Football and Basketball Success on Graduation Rates and Alumni Giving Rates," *Economics of Education Review* 23, 655–661.

Vella, F., and M. Verbeek (1998), "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," *Journal of Applied Econometrics* 13, 163–183.

Wald, A. (1940), "The Fitting of Straight Lines If Both Variables Are Subject to Error," *Annals of Mathematical Statistics* 11, 284–300.

Wallis, K. F. (1972), "Testing for Fourth-Order Autocorrelation in Quarterly Regression Equations," *Econometrica* 40, 617–636.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48, 817–838.

White, H. (1984), *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

White, M. J. (1986), "Property Taxes and Firm Location: Evidence from Proposition 13," in *Studies in State and Local Public Finance*, ed. H. S. Rosen, 83–112. Chicago: University of Chicago Press.

Whittington, L. A., J. Alm, and H. E. Peters (1990), "Fertility and the Personal Exemption: Implicit Pronatalist Policy in the United States," *American Economic Review* 80, 545–556.

Wooldridge, J. M. (1989), "A Computationally Simple Heteroskedasticity and Serial Correlation-Robust Standard Error for the Linear Regression Model," *Economics Letters* 31, 239–243.

Wooldridge, J. M. (1991a), "A Note on Computing R-Squared and Adjusted R-Squared for Trending and Seasonal Data," *Economics Letters* 36, 49–54.

Wooldridge, J. M. (1991b), "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances," *Journal of Econometrics* 47, 5–46.

Wooldridge, J. M. (1994a), "A Simple Specification Test for the Predictive Ability of Transformation Models," *Review of Economics and Statistics* 76, 59–65.

Wooldridge, J. M. (1994b), "Estimation and Inference for Dependent Processes," in *Handbook of Econometrics*, volume 4, chapter 45, ed. R. F. Engle and D. L. McFadden, 2639–2738. Amsterdam: North-Holland.

Wooldridge, J. M. (1995), "Score Diagnostics for Linear Models Estimated by Two Stage Least Squares," in *Advances in Econometrics and Quantitative Economics*, ed. G. S. Maddala, P. C. B. Phillips, and T. N. Srinivasan, 66–87. Oxford: Blackwell.

Wooldridge, J. M. (2001), "Diagnostic Testing," in *Companion to Theoretical Econometrics*, ed. B. H. Baltagi, 180–200. Oxford: Blackwell.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

Glossary

A

Adjusted *R*-Squared: A goodness-of-fit measure in multiple regression analysis that penalizes additional explanatory variables by using a degrees of freedom adjustment in estimating the error variance.

Alternative Hypothesis: The hypothesis against which the null hypothesis is tested.

AR(1) Serial Correlation: The errors in a time series regression model follow an AR(1) model.

Asymptotic Bias: *See* inconsistency.

Asymptotic Confidence Interval: A confidence interval that is approximately valid in large sample sizes.

Asymptotic Normality: The sampling distribution of a properly normalized estimator converges to the standard normal distribution.

Asymptotic Properties: Properties of estimators and test statistics that apply when the sample size grows without bound.

Asymptotic Standard Error: A standard error that is valid in large samples.

Asymptotic *t* Statistic: A *t* statistic that has an approximate standard normal distribution in large samples.

Asymptotic Variance: The square of the value by which we must divide an estimator in order to obtain an asymptotic standard normal distribution.

Asymptotically Efficient: For consistent estimators with asymptotically normal distributions, the estimator with the smallest asymptotic variance.

Asymptotically Uncorrelated: A time series process in which the correlation between random variables at two points in time tends to zero as the time interval between them increases. (*See also* weakly dependent.)

Attenuation Bias: Bias in an estimator that is always toward zero; thus, the expected value of an estimator with attenuation bias is less in magnitude than the absolute value of the parameter.

Augmented Dickey-Fuller Test: A test for a unit root that includes lagged changes of the variable as regressors.

Autocorrelation: *See* serial correlation.

Autoregressive Conditional Heteroskedasticity (ARCH): A model of dynamic heteroskedasticity where the variance of the error term, given past information, depends linearly on the past squared errors.

Autoregressive Process of Order One [AR(1)]: A time series model whose current value depends linearly on its most recent value plus an unpredictable disturbance.

Auxiliary Regression: A regression used to compute a test statistic—such as the test statistics for heteroskedasticity and serial correlation—or any other regression that does not estimate the model of primary interest.

Average Marginal Effect: *See* average partial effect.

Average Partial Effect (APE): For nonconstant partial effects, the partial effect averaged across the specified population.

Average Causal Effect (ACE): *See* average treatment effect.

Average Treatment Effect (ATE): A treatment, or policy, effect averaged across the population.

B

Balanced Panel: A panel data set where all years (or periods) of data are available for all cross-sectional units.

Base Group: The group represented by the overall intercept in a multiple regression model that includes dummy explanatory variables.

Base Period: For index numbers, such as price or production indices, the period against which all other time periods are measured.

Base Value: The value assigned to the base period for constructing an index number; usually the base value is 1 or 100.

Benchmark Group: *See* base group.

Best Linear Unbiased Estimator (BLUE): Among all linear unbiased estimators, the one with the smallest variance. OLS is BLUE, conditional on the sample values of the explanatory variables, under the Gauss-Markov assumptions.

Beta Coefficients: *See* standardized coefficients.

Bias: The difference between the expected value of an estimator and the population value that the estimator is supposed to be estimating.

Biased Estimator: An estimator whose expectation, or sampling mean, is different from the population value it is supposed to be estimating.

Biased Toward Zero: A description of an estimator whose expectation in absolute value is less than the absolute value of the population parameter.

Binary (Dummy) Variable: *See* dummy variable.

Binary Response Model: A model for a binary (dummy) dependent variable.

Binary Variable: *See* dummy variable.

Binomial Distribution: The probability distribution of the number of successes out of n independent Bernoulli trials, where each trial has the same probability of success.

BLUE: *See* best linear unbiased estimator.

Bootstrap: A resampling method that draws random samples, with replacement, from the original data set.

Bootstrap Standard Error: A standard error obtained as the sample standard deviation of an estimate across all bootstrap samples.

Breusch-Godfrey Test: An asymptotically justified test for AR(p) serial correlation, with AR(1) being the most popular; the test allows for lagged dependent variables as well as other regressors that are not strictly exogenous.

Breusch-Pagan Test for Heteroskedasticity (BP Test): Refer to Breusch-Pegan Test.

C

Causal Effect: A ceteris paribus change in one variable that has an effect on another variable.

Causal (Treatment) Effect: The difference in outcomes between when an observation has a treatment (e.g. policy) and when it is not treated.

Censored Normal Regression Model: The special case of the censored regression model where the underlying population model satisfies the classical linear model assumptions.

Censored Regression Model: A multiple regression model where the dependent variable has been censored above or below some known threshold.

Central Limit Theorem (CLT): A key result from probability theory which implies that the sum of independent random variables, or even weakly dependent random variables, when standardized by its standard deviation, has a distribution that tends to standard normal as the sample size grows.

Ceteris Paribus: All other relevant factors are held fixed.

Chi-Square Distribution: A probability distribution obtained by adding the squares of independent standard normal random variables. The number of terms in the sum equals the degrees of freedom in the distribution.

Chi-Square Random Variable: A random variable with a chi-square distribution.

Chow Statistic: An F statistic for testing the equality of regression parameters across different groups (say, men and women) or time periods (say, before and after a policy change).

Classical Errors-in-Variables (CEV): A measurement error model where the observed measure equals the actual variable plus an independent, or at least an uncorrelated, measurement error.

Classical Linear Model: The multiple linear regression model under the full set of classical linear model assumptions.

Classical Linear Model (CLM) Assumptions: The ideal set of assumptions for multiple regression analysis: for cross-sectional analysis, Assumptions MLR.1 through MLR.6, and for time series analysis, Assumptions TS.1 through TS.6. The assumptions include linearity in the parameters, no perfect collinearity, the zero conditional mean assumption, homoskedasticity, no serial correlation, and normality of the errors.

Cluster Effect: An unobserved effect that is common to all units, usually people, in the cluster.

Cluster-Robust Standard Errors: Standard error estimates that allow for unrestricted forms of serial correlation and heteroskedasticity in panel data. These standard errors require a large cross section (N) and not too large time series (T).

Cluster Sample: A sample of natural clusters or groups that usually consist of people.

Clustering: The act of computing standard errors and test statistics that are robust to cluster correlation, either due to cluster sampling or to time series correlation in panel data.

Cochrane-Orcutt (CO) Estimation: A method of estimating a multiple linear regression model with AR(1) errors and strictly exogenous explanatory variables; unlike Prais-Winsten, Cochrane-Orcutt does not use the equation for the first time period.

Coefficient of Determination: *See* R-squared.

Cointegration: The notion that a linear combination of two series, each of which is integrated of order one, is integrated of order zero.

Column Vector: A vector of numbers arranged as a column.

Complete Cases Indicator: A dummy variable that is equal to 1 if and only if we have data for all variables for a particular observation and 0 otherwise.

Composite Error: Refer to Composite Error Term.

Composite Error Term: In a panel data model, the sum of the time-constant unobserved effect and the idiosyncratic error.

Conditional Distribution: The probability distribution of one random variable, given the values of one or more other random variables.

Conditional Expectation: The expected or average value of one random variable, called the dependent or explained

variable, that depends on the values of one or more other variables, called the independent or explanatory variables.

Conditional Forecast: A forecast that assumes the future values of some explanatory variables are known with certainty.

Conditional Independence: When treatment and outcome variables can be considered to be independent of one another after conditioning on control variables.

Conditional Median: The median of a response variable conditional on some explanatory variables.

Conditional Variance: The variance of one random variable, given one or more other random variables.

Confidence Interval (CI): A rule used to construct a random interval so that a certain percentage of all data sets, determined by the confidence level, yields an interval that contains the population value.

Consistency: An estimator converges in probability to the correct population value as the sample size grows.

Consistent Estimator: An estimator that converges in probability to the population parameter as the sample size grows without bound.

Consistent Test: A test where, under the alternative hypothesis, the probability of rejecting the null hypothesis converges to one as the sample size grows without bound.

Constant Elasticity Model: A model where the elasticity of the dependent variable, with respect to an explanatory variable, is constant; in multiple regression, both variables appear in logarithmic form.

Contemporaneously Homoskedastic: Describes a time series or panel data applications in which the variance of the error term, conditional on the regressors in the same time period, is constant.

Contemporaneously Exogenous: Describes a time series or panel data application in which a regressor is contemporaneously exogenous if it is uncorrelated with the error term in the same time period, although it may be correlated with the errors in other time periods.

Continuous Random Variable: A random variable that takes on any particular value with probability zero.

Control Group: In program evaluation, the group that does not participate in the program.

Control Variable: *See* explanatory variable.

Corner Solution Response: A nonnegative dependent variable that is roughly continuous over strictly positive values but takes on the value zero with some regularity.

Correlated Random Effects: An approach to panel data analysis where the correlation between the unobserved effect and the explanatory variables is modeled, usually as a linear relationship.

Correlation Coefficient: A measure of linear dependence between two random variables that does not depend on units of measurement and is bounded between −1 and 1.

Count Variable: A variable that takes on nonnegative integer values.

Counterfactual Outcomes: The different outcomes that result from a counterfactual reasoning process.

Counterfactual Reasonings: A method of policy evaluation in which we imagine an identical observation (individual, firm, country, etc.) under two different states of the world (e.g. with a policy and without a policy).

Covariance: A measure of linear dependence between two random variables.

Covariance Stationary: A time series process with constant mean and variance where the covariance between any two random variables in the sequence depends only on the distance between them.

Covariate: *See* explanatory variable.

Critical Value: In hypothesis testing, the value against which a test statistic is compared to determine whether or not the null hypothesis is rejected.

Cross-Sectional Data Set: A data set collected by sampling a population at a given point in time.

Cumulative Distribution Function (cdf): A function that gives the probability of a random variable being less than or equal to any specified real number.

Cumulative Effect: At any point in time, the change in a response variable after a permanent increase in an explanatory variable—usually in the context of distributed lag models.

D

Data Frequency: The interval at which time series data are collected. Yearly, quarterly, and monthly are the most common data frequencies.

Data Mining: The practice of using the same data set to estimate numerous models in a search to find the "best" model.

Davidson-MacKinnon Test: A test that is used for testing a model against a nonnested alternative; it can be implemented as a t test on the fitted values from the competing model.

Degrees of Freedom (df): In multiple regression analysis, the number of observations minus the number of estimated parameters.

Denominator Degrees of Freedom: In an F test, the degrees of freedom in the unrestricted model.

Dependent Variable: The variable to be explained in a multiple regression model (and a variety of other models).

Derivative: The slope of a smooth function, as defined using calculus.

Descriptive Statistic: A statistic used to summarize a set of numbers; the sample average, sample median, and sample standard deviation are the most common.

Deseasonalizing: The removing of the seasonal components from a monthly or quarterly time series.

Detrending: The practice of removing the trend from a time series.

Diagonal Matrix: A matrix with zeros for all off-diagonal entries.

Dickey-Fuller Distribution: The limiting distribution of the *t* statistic in testing the null hypothesis of a unit root.

Dickey-Fuller (DF) Test: A *t* test of the unit root null hypothesis in an AR(1) model. (*See also* augmented Dickey-Fuller test.)

Difference in Slopes: A description of a model where some slope parameters may differ by group or time period.

Difference-in-Differences (DD or DID) Estimator: An estimator that arises in policy analysis with data for two time periods. One version of the estimator applies to independently pooled cross sections and another to panel data sets.

Difference-in-Difference-in-Differences (DDD) Estimator: An estimator that allows for one additional control group than the standard difference-in-differences estimator. Useful in dealing with violations of the parallel trends assumption.

Difference-Stationary Process: A time series sequence that is I(0) in its first differences.

Discrete Random Variable: A random variable that takes on at most a finite or countably infinite number of values.

Disturbance: *See* error term.

Downward Bias: The expected value of an estimator is below the population value of the parameter.

Dummy Dependent Variable: *See* binary response model.

Dummy Variable: A variable that takes on the value zero or one.

Dummy Variable Regression: In a panel data setting, the regression that includes a dummy variable for each cross-sectional unit, along with the remaining explanatory variables. It produces the fixed effects estimator.

Dummy Variable Trap: The mistake of including too many dummy variables among the independent variables; it occurs when an overall intercept is in the model and a dummy variable is included for each group.

Duration Analysis: An application of the censored regression model where the dependent variable is time elapsed until a certain event occurs, such as the time before an unemployed person becomes reemployed.

Durbin-Watson (*DW*) Statistic: A statistic used to test for first order serial correlation in the errors of a time series regression model under the classical linear model assumptions.

Dynamically Complete Model: A time series model where no further lags of either the dependent variable or the explanatory variables help to explain the mean of the dependent variable.

E

Econometric Model: An equation relating the dependent variable to a set of explanatory variables and unobserved disturbances, where unknown population parameters determine the ceteris paribus effect of each explanatory variable.

Economic Model: A relationship derived from economic theory or less formal economic reasoning.

Economic Significance: *See* practical significance.

Elasticity: The percentage change in one variable given a 1% ceteris paribus increase in another variable.

Empirical Analysis: A study that uses data in a formal econometric analysis to test a theory, estimate a relationship, or determine the effectiveness of a policy.

Endogenous Explanatory Variable: An explanatory variable in a multiple regression model that is correlated with the error term, either because of an omitted variable, measurement error, or simultaneity.

Endogenous Sample Selection: Nonrandom sample selection where the selection is related to the dependent variable, either directly or through the error term in the equation.

Endogenous Variables: In simultaneous equations models, variables that are determined by the equations in the system.

Engle-Granger Test: A test of the null hypothesis that two time series are not cointegrated; the statistic is obtained as the Dickey-Fuller statistic using OLS residuals.

Engle-Granger Two-Step Procedure: A two-step method for estimating error correction models whereby the cointegrating parameter is estimated in the first stage, and the error correction parameters are estimated in the second.

Error Correction Model: A time series model in first differences that also contains an error correction term, which works to bring two I(1) series back into long-run equilibrium.

Error Term (Disturbance): The variable in a simple or multiple regression equation that contains unobserved factors which affect the dependent variable. The error term may also include measurement errors in the observed dependent or independent variables.

Error Variance: The variance of the error term in a multiple regression model.

Errors-in-Variables: A situation where either the dependent variable or some independent variables are measured with error.

Estimate: The numerical value taken on by an estimator for a particular sample of data.

Estimator: A rule for combining data to produce a numerical value for a population parameter; the form of the rule does not depend on the particular sample obtained.

Event Study: An econometric analysis of the effects of an event, such as a change in government regulation or economic policy, on an outcome variable.

Excluding a Relevant Variable: In multiple regression analysis, leaving out a variable that has a nonzero partial effect on the dependent variable.

Exclusion Restrictions: Restrictions which state that certain variables are excluded from the model (or have zero population coefficients).

Exogenous Explanatory Variable: An explanatory variable that is uncorrelated with the error term.

Exogenous Sample Selection: A sample selection that either depends on exogenous explanatory variables or is independent of the error term in the equation of interest.

Exogenous Variable: Any variable that is uncorrelated with the error term in the model of interest.

Expected Value: A measure of central tendency in the distribution of a random variable, including an estimator.

Experiment: In probability, a general term used to denote an event whose outcome is uncertain. In econometric analysis, it denotes a situation where data are collected by randomly assigning individuals to control and treatment groups.

Experimental Data: Data that have been obtained by running a controlled experiment.

Experimental Group: *See* treatment group.

Explained Sum of Squares (SSE): The total sample variation of the fitted values in a multiple regression model.

Explained Variable: *See* dependent variable.

Explanatory Variable: In regression analysis, a variable that is used to explain variation in the dependent variable.

Exponential Function: A mathematical function defined for all values that has an increasing slope but a constant proportionate change.

Exponential Smoothing: A simple method of forecasting a variable that involves a weighting of all previous outcomes on that variable.

Exponential Trend: A trend with a constant growth rate.

F

***F* Distribution:** The probability distribution obtained by forming the ratio of two independent chi-square random variables, where each has been divided by its degrees of freedom.

***F* Random Variable:** A random variable with an *F* distribution.

***F* Statistic:** A statistic used to test multiple hypotheses about the parameters in a multiple regression model.

Falsification test: A method of testing the strict exogeneity assumption that includes a future value of a policy variable as a determinant of the current value of the outcome variable.

Feasible GLS (FGLS) Estimator: A GLS procedure where variance or correlation parameters are unknown and therefore must first be estimated. (*See also* generalized least squares estimator.)

Finite Distributed Lag (FDL) Model: A dynamic model where one or more explanatory variables are allowed to have lagged effects on the dependent variable.

First Difference: A transformation on a time series constructed by taking the difference of adjacent time periods, where the earlier time period is subtracted from the later time period.

First-Differenced Equation: In time series or panel data models, an equation where the dependent and independent variables have all been first differenced.

First-Differenced Estimator: In a panel data setting, the pooled OLS estimator applied to first differences of the data across time.

First Order Autocorrelation: For a time series process ordered chronologically, the correlation coefficient between pairs of adjacent observations.

First Order Conditions: The set of linear equations used to solve for the OLS estimates.

First Stage: The first stage of a 2SLS procedure in which the endogenous explanatory variable is regressed on all instruments and exogenous explanatory variables.

Fitted Values: The estimated values of the dependent variable when the values of the independent variables for each observation are plugged into the OLS regression line.

Fixed Effect: *See* unobserved effect.

Fixed Effects Estimator: For the unobserved effects panel data model, the estimator obtained by applying pooled OLS to a time-demeaned equation.

Fixed Effects Model: An unobserved effects panel data model where the unobserved effects are allowed to be arbitrarily correlated with the explanatory variables in each time period.

Fixed Effects Transformation: For panel data, the time-demeaned data.

Forecast Error: The difference between the actual outcome and the forecast of the outcome.

Forecast Interval: In forecasting, a confidence interval for a yet unrealized future value of a time series variable. (*See also* prediction interval.)

Frisch-Waugh Theorem: The general algebraic result that provides multiple regression analysis with its "partialling out" interpretation.

Functional Form Misspecification: A problem that occurs when a model has omitted functions of the explanatory variables (such as quadratics) or uses the wrong functions of either the dependent variable or some explanatory variables.

G

Gauss-Markov Assumptions: The set of assumptions (Assumptions MLR.1 through MLR.5 or TS.1 through TS.5) under which OLS is BLUE.

Gauss-Markov Theorem: The theorem that states that, under the five Gauss-Markov assumptions (for cross-sectional or time series models), the OLS estimator is BLUE (conditional on the sample values of the explanatory variables).

Generalized Least Squares (GLS) Estimator: An estimator that accounts for a known structure of the error variance (heteroskedasticity), serial correlation pattern in the errors, or both, via a transformation of the original model.

Geometric (or Koyck) Distributed Lag: An infinite distributed lag model where the lag coefficients decline at a geometric rate.

Granger Causality: A limited notion of causality where past values of one series (x_t) are useful for predicting future values of another series (y_t), after past values of y_t have been controlled for.

Group-Specific: Time trends in panel data that are allowed to vary by group (as opposed to imposing a common time trend for all observations).

Growth Rate: The proportionate change in a time series from the previous period. It may be approximated as the difference in logs or reported in percentage form.

H

Heckit Method: An econometric procedure used to correct for sample selection bias due to incidental truncation or some other form of nonrandomly missing data.

Heterogeneity Bias: The bias in OLS due to omitted heterogeneity (or omitted variables).

Heterogeneous trend model: A panel data model that allows the time trend to vary across individual observations. The model is estimated in first differences and requires at least three time periods of data.

Heteroskedasticity: The variance of the error term, given the explanatory variables, is not constant.

Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors: A form of the OLS standard errors that is robust to both heteroskedasticity and serial correlation.

Heteroskedasticity of Unknown Form: Heteroskedasticity that may depend on the explanatory variables in an unknown, arbitrary fashion.

Heteroskedasticity-Robust *F* Statistic: An *F*-type statistic that is (asymptotically) robust to heteroskedasticity of unknown form.

Heteroskedasticity-Robust *LM* Statistic: An *LM* statistic that is robust to heteroskedasticity of unknown form.

Heteroskedasticity-Robust Standard Error: A standard error that is (asymptotically) robust to heteroskedasticity of unknown form.

Heteroskedasticity-Robust *t* Statistic: A *t* statistic that is (asymptotically) robust to heteroskedasticity of unknown form.

Highly Persistent: A time series process where outcomes in the distant future are highly correlated with current outcomes.

Homoskedasticity: The errors in a regression model have constant variance conditional on the explanatory variables.

Hypothesis Test: A statistical test of the null, or maintained, hypothesis against an alternative hypothesis.

I

Idempotent Matrix: A (square) matrix where multiplication of the matrix by itself equals itself.

Identification: A population parameter, or set of parameters, can be consistently estimated.

Identified Equation: An equation whose parameters can be consistently estimated, especially in models with endogenous explanatory variables.

Identity Matrix: A square matrix where all diagonal elements are one and all off-diagonal elements are zero.

Idiosyncratic Error: In panel data models, the error that changes over time as well as across units (say, individuals, firms, or cities).

Ignorable Assignment: *See* conditional independence.

Impact Elasticity: In a distributed lag model, the immediate percentage change in the dependent variable given a 1% increase in the independent variable.

Impact Multiplier: *See* impact propensity.

Impact Propensity: In a distributed lag model, the immediate change in the dependent variable given a one-unit increase in the independent variable.

Incidental Truncation: A sample selection problem whereby one variable, usually the dependent variable, is only observed for certain outcomes of another variable.

Inclusion of an Irrelevant Variable: The including of an explanatory variable in a regression model that has a zero population parameter in estimating an equation by OLS.

Inconsistency: The difference between the probability limit of an estimator and the parameter value.

Inconsistent: Describes an estimator that does not converge (in probability) to the correct population parameter as the sample size grows.

Independent Random Variables: Random variables whose joint distribution is the product of the marginal distributions.

Independent Variable: *See* explanatory variable.

Independently Pooled Cross Section: A data set obtained by pooling independent random samples from different points in time.

Index Number: A statistic that aggregates information on economic activity, such as production or prices.

Infinite Distributed Lag (IDL) Model: A distributed lag model where a change in the explanatory variable can have an impact on the dependent variable into the indefinite future.

Influential Observations: *See* outliers.

Information Set: In forecasting, the set of variables that we can observe prior to forming our forecast.

In-Sample Criteria: Criteria for choosing forecasting models that are based on goodness-of-fit within the sample used to obtain the parameter estimates.

Instrument: *See* instrumental variable.

Instrument Exogeneity: In instrumental variables estimation, the requirement that an instrumental variable is uncorrelated with the error term.

Instrument Relevance: In instrumental variables estimation, the requirement that an instrumental variable helps to partially explain variation in the endogenous explanatory variable.

Instrumental Variable: In an equation with an endogenous explanatory variable, an IV is a variable that does not appear in the equation, is uncorrelated with the error in the equation, and is (partially) correlated with the endogenous explanatory variable.

Instrumental Variables (IV) Estimator: An estimator in a linear model used when instrumental variables are available for one or more endogenous explanatory variables.

Integrated of Order One [I(1)]: A time series process that needs to be first-differenced in order to produce an I(0) process.

Integrated of Order Zero [I(0)]: A stationary, weakly dependent time series process that, when used in regression analysis, satisfies the law of large numbers and the central limit theorem.

Interaction Effect: In multiple regression, the partial effect of one explanatory variable depends on the value of a different explanatory variable.

Interaction Term: An independent variable in a regression model that is the product of two explanatory variables.

Intercept: In the equation of a line, the value of the y variable when the x variable is zero.

Intercept Parameter: The parameter in a multiple linear regression model that gives the expected value of the dependent variable when all the independent variables equal zero.

Intercept Shift: The intercept in a regression model differs by group or time period.

Internet: A global computer network that can be used to access information and download databases.

Interval Estimator: A rule that uses data to obtain lower and upper bounds for a population parameter. (*See also* confidence interval.)

Inverse: For an $n \times n$ matrix, its inverse (if it exists) is the $n \times n$ matrix for which pre- and post-multiplication by the original matrix yields the identity matrix.

Inverse Mills Ratio: A term that can be added to a multiple regression model to remove sample selection bias.

J

Joint Distribution: The probability distribution determining the probabilities of outcomes involving two or more random variables.

Joint Hypotheses Test: A test involving more than one restriction on the parameters in a model.

Jointly Insignificant: Failure to reject, using an F test at a specified significance level, that all coefficients for a group of explanatory variables are zero.

Jointly Statistically Significant: The null hypothesis that two or more explanatory variables have zero population coefficients is rejected at the chosen significance level.

Just Identified Equation: For models with endogenous explanatory variables, an equation that is identified but would not be identified with one fewer instrumental variable.

K

Kurtosis: A measure of the thickness of the tails of a distribution based on the fourth moment of the standardized random variable; the measure is usually compared to the value for the standard normal distribution, which is three.

L

Lag Distribution: In a finite or infinite distributed lag model, the lag coefficients graphed as a function of the lag length.

Lagged Dependent Variable: An explanatory variable that is equal to the dependent variable from an earlier time period.

Lagged Endogenous Variable: In a simultaneous equations model, a lagged value of one of the endogenous variables.

Lagrange Multiplier (LM) Statistic: A test statistic with large-sample justification that can be used to test for omitted variables, heteroskedasticity, and serial correlation, among other model specification problems.

Large Sample Properties: *See* asymptotic properties.

Latent Variable Model: A model where the observed dependent variable is assumed to be a function of an underlying latent, or unobserved, variable.

Law of Iterated Expectations: A result from probability that relates unconditional and conditional expectations.

Law of Large Numbers (LLN): A theorem that says that the average from a random sample converges in probability to the population average; the LLN also holds for stationary and weakly dependent time series.

Leads and Lags Estimator: An estimator of a cointegrating parameter in a regression with I(1) variables, where the current, some past, and some future first differences in the explanatory variable are included as regressors.

Least Absolute Deviations (LAD): A method for estimating the parameters of a multiple regression model based on minimizing the sum of the absolute values of the residuals.

Least Squares Estimator: An estimator that minimizes a sum of squared residuals.

Likelihood Ratio Statistic: A statistic that can be used to test single or multiple hypotheses when the constrained and unconstrained models have been estimated by maximum likelihood. The statistic is twice the difference in the unconstrained and constrained log-likelihoods.

Limited Dependent Variable (LDV): A dependent or response variable whose range is restricted in some important way.

Linear Function: A function where the change in the dependent variable, given a one-unit change in an independent variable, is constant.

Linear Probability Model (LPM): A binary response model where the response probability is linear in its parameters.

Linear Time Trend: A trend that is a linear function of time.

Linearly Independent Vectors: A set of vectors such that no vector can be written as a linear combination of the others in the set.

Log Function: A mathematical function, defined only for strictly positive arguments, with a positive but decreasing slope.

Logit Model: A model for binary response where the response probability is the logit function evaluated at a linear function of the explanatory variables.

Log-Likelihood Function: The sum of the log-likelihoods, where the log-likelihood for each observation is the log of the density of the dependent variable given the explanatory variables; the log-likelihood function is viewed as a function of the parameters to be estimated.

Longitudinal Data: *See* panel data.

Long-Run Elasticity: The long-run propensity in a distributed lag model with the dependent and independent variables in logarithmic form; thus, the long-run elasticity is the eventual percentage increase in the explained variable, given a permanent 1% increase in the explanatory variable.

Long-Run Multiplier: *See* long-run propensity.

Long-Run Propensity (LRP): In a distributed lag model, the eventual change in the dependent variable given a permanent, one-unit increase in the independent variable.

Loss Function: A function that measures the loss when a forecast differs from the actual outcome; the most common examples are absolute value loss and squared loss.

M

Marginal Effect: The effect on the dependent variable that results from changing an independent variable by a small amount.

Martingale: A time series process whose expected value, given all past outcomes on the series, simply equals the most recent value.

Martingale Difference Sequence: The first difference of a martingale. It is unpredictable (or has a zero mean), given past values of the sequence.

Matched Pair Sample: A sample where each observation is matched with another, as in a sample consisting of a husband and wife or a set of two siblings.

Matrix: An array of numbers.

Matrix Multiplication: An algorithm for multiplying together two conformable matrices.

Matrix Notation: A convenient mathematical notation, grounded in matrix algebra, for expressing and manipulating the multiple regression model.

Maximum Likelihood Estimation (MLE): A broadly applicable estimation method where the parameter estimates are chosen to maximize the log-likelihood function.

Maximum Likelihood Estimator: An estimator that maximizes the (log of the) likelihood function.

Mean Absolute Error (MAE): A performance measure in forecasting, computed as the average of the absolute values of the forecast errors.

Mean Independent: The key requirement in multiple regression analysis, which says the unobserved error has a mean that does not change across subsets of the population defined by different values of the explanatory variables.

Mean Squared Error (MSE): The expected squared distance that an estimator is from the population value; it equals the variance plus the square of any bias.

Measurement Error: The difference between an observed variable and the variable that belongs in a multiple regression equation.

Median: In a probability distribution, it is the value where there is a 50% chance of being below the value and a 50% chance of being above it. In a sample of numbers, it is the middle value after the numbers have been ordered.

Micronumerosity: A term introduced by Arthur Goldberger to describe properties of econometric estimators with small sample sizes.

Minimum Variance Unbiased Estimator: An estimator with the smallest variance in the class of all unbiased estimators.

Missing at Random: In multiple regression analysis, a missing data mechanism where the reason data are missing may be correlated with the explanatory variables but is independent of the error term.

Missing Completely at Random (MCAR): In multiple regression analysis, a missing data mechanism where the reason data are missing is statistically independent of the values of the explanatory variables as well as the unobserved error.

Missing Data: A data problem that occurs when we do not observe values on some variables for certain observations (individuals, cities, time periods, and so on) in the sample.

Missing Indicator Method: A method for dealing with missing observations in an explanatory variable. The explanatory variable is included alongside a binary variable equal to 0 when the explanatory variable is missing for that observation, allowing us to use the full data set.

Misspecification Analysis: The process of determining likely biases that can arise from omitted variables, measurement error, simultaneity, and other kinds of model misspecification.

Moving Average Process of Order One [MA(1)]: A time series process generated as a linear function of the current value and one lagged value of a zero-mean, constant variance, uncorrelated stochastic process.

Multicollinearity: A term that refers to correlation among the independent variables in a multiple regression model; it is usually invoked when some correlations are "large," but an actual magnitude is not well defined.

Multiple Hypotheses Test: A test of a null hypothesis involving more than one restriction on the parameters.

Multiple Linear Regression (MLR) Model: A model linear in its parameters, where the dependent variable is a function of independent variables plus an error term.

Multiple Regression Analysis: A type of analysis that is used to describe estimation of and inference in the multiple linear regression model.

Multiple Restrictions: More than one restriction on the parameters in an econometric model.

Multiple-Step-Ahead Forecast: A time series forecast of more than one period into the future.

Multiplicative Measurement Error: Measurement error where the observed variable is the product of the true unobserved variable and a positive measurement error.

Multivariate Normal Distribution: A distribution for multiple random variables where each linear combination of the random variables has a univariate (one-dimensional) normal distribution.

N

***n*-*R*-Squared Statistic:** *See* Lagrange multiplier statistic.

Natural Experiment: A situation where the economic environment—sometimes summarized by an explanatory variable—exogenously changes, perhaps inadvertently, due to a policy or institutional change.

Natural Logarithm: *See* logarithmic function.

Newey-West Standards Errors: A specific form of HAC standard errors. In this case, the truncation lag is set to the integer part of 4(n/100) superscript 2/9. Refer to file submitted with gloss file.

Nonexperimental Data: Data that have not been obtained through a controlled experiment.

Nonlinear Function: A function whose slope is not constant.

Nonnested Models: Two (or more) models where no model can be written as a special case of the other by imposing restrictions on the parameters.

Nonrandom Sample: A sample obtained other than by sampling randomly from the population of interest.

Nonrandom Sample Selection: When the sample is not randomly drawn from the population, but is selected on the basis of individual characteristics.

Nonstationary Process: A time series process whose joint distributions are not constant across different epochs.

Normal Distribution: A probability distribution commonly used in statistics and econometrics for modeling a population. Its probability distribution function has a bell shape.

Normality Assumption: The classical linear model assumption which states that the error (or dependent variable) has a normal distribution, conditional on the explanatory variables.

Null Hypothesis: In classical hypothesis testing, we take this hypothesis as true and require the data to provide substantial evidence against it.

Numerator Degrees of Freedom: In an F test, the number of restrictions being tested.

O

Observational Data: *See* nonexperimental data.

OLS: *See* ordinary least squares.

OLS Intercept Estimate: The intercept in an OLS regression line.

OLS Regression Line: The equation relating the predicted value of the dependent variable to the independent variables, where the parameter estimates have been obtained by OLS.

OLS Slope Estimate: A slope in an OLS regression line.

Omitted Variable Bias: The bias that arises in the OLS estimators when a relevant variable is omitted from the regression.

Omitted Variables: One or more variables, which we would like to control for, have been omitted in estimating a regression model.

One-Sided Alternative: An alternative hypothesis that states that the parameter is greater than (or less than) the value hypothesized under the null.

One-Step-Ahead Forecast: A time series forecast one period into the future.

One-Tailed Test: A hypothesis test against a one-sided alternative.

Online Databases: Databases that can be accessed via a computer network.

Online Search Services: Computer software that allows the Internet or databases on the Internet to be searched by topic, name, title, or keywords.

Order Condition: A necessary condition for identifying the parameters in a model with one or more endogenous explanatory variables: the total number of exogenous variables must be at least as great as the total number of explanatory variables.

Ordinal Variable: A variable where the ordering of the values conveys information but the magnitude of the values does not.

Ordinary Least Squares (OLS): A method for estimating the parameters of a multiple linear regression model. The ordinary least squares estimates are obtained by minimizing the sum of squared residuals.

Outliers: Observations in a data set that are substantially different from the bulk of the data, perhaps because of errors or because some data are generated by a different model than most of the other data.

Out-of-Sample Criteria: Criteria used for choosing forecasting models which are based on a part of the sample that was not used in obtaining parameter estimates.

Over Controlling: In a multiple regression model, including explanatory variables that should not be held fixed when studying the ceteris paribus effect of one or more other explanatory variables; this can occur when variables that are themselves outcomes of an intervention or a policy are included among the regressors.

Overall Significance of a Regression: A test of the joint significance of all explanatory variables appearing in a multiple regression equation.

Overdispersion: In modeling a count variable, the variance is larger than the mean.

Overidentified Equation: In models with endogenous explanatory variables, an equation where the number of instrumental variables is strictly greater than the number of endogenous explanatory variables.

Overidentifying Restrictions: The extra moment conditions that come from having more instrumental variables than endogenous explanatory variables in a linear model.

Overspecifying the Model: *See* inclusion of an irrelevant variable.

P

***p*-Value:** The smallest significance level at which the null hypothesis can be rejected. Equivalently, the largest significance level at which the null hypothesis cannot be rejected.

Pairwise Uncorrelated Random Variables: A set of two or more random variables where each pair is uncorrelated.

Panel Data: A data set constructed from repeated cross sections over time. With a *balanced* panel, the same units appear in each time period. With an *unbalanced* panel, some units do not appear in each time period, often due to attrition.

Parallel Trends Assumption: The assumption that any trends in the outcome variable would trend at the same rate and direction between the treatment and control groups in the absence of the treatment.

Partial Derivative: For a smooth function of more than one variable, the slope of the function in one direction.

Partial Effect: The effect of an explanatory variable on the dependent variable, holding other factors in the regression model fixed.

Partial Effect at the Average (PEA): In models with nonconstant partial effects, the partial effect evaluated at the average values of the explanatory variables.

Percent Correctly Predicted: In a binary response model, the percentage of times the prediction of zero or one coincides with the actual outcome.

Percentage Change: The proportionate change in a variable, multiplied by 100.

Percentage Point Change: The change in a variable that is measured as a percentage.

Perfect Collinearity: In multiple regression, one independent variable is an exact linear function of one or more other independent variables.

Plug-In Solution to the Omitted Variables Problem: A proxy variable is substituted for an unobserved omitted variable in an OLS regression.

Point Forecast: The forecasted value of a future outcome.

Poisson Distribution: A probability distribution for count variables.

Poisson Regression Model: A model for a count dependent variable where the dependent variable, conditional on the explanatory variables, is nominally assumed to have a Poisson distribution.

Policy Analysis: An empirical analysis that uses econometric methods to evaluate the effects of a certain policy.

Pooled Cross Section: A data configuration where independent cross sections, usually collected at different points in time, are combined to produce a single data set.

Population Model: A model, especially a multiple linear regression model, that describes a population.

Population *R*-Squared: In the population, the fraction of the variation in the dependent variable that is explained by the explanatory variables.

Population Regression Function: *See* conditional expectation.

Positive Definite: A symmetric matrix such that all quadratic forms, except the trivial one that must be zero, are strictly positive.

Positive Semi-Definite: A symmetric matrix such that all quadratic forms are nonnegative.

Power of a Test: The probability of rejecting the null hypothesis when it is false; the power depends on the values of the population parameters under the alternative.

Practical Significance: The practical or economic importance of an estimate, which is measured by its sign and magnitude, as opposed to its statistical significance.

Prais-Winsten (PW) Estimation: A method of estimating a multiple linear regression model with AR(1) errors and strictly exogenous explanatory variables; unlike Cochrane-Orcutt, Prais-Winsten uses the equation for the first time period in estimation.

Predetermined Variable: In a simultaneous equations model, either a lagged endogenous variable or a lagged exogenous variable.

Predicted Variable: *See* dependent variable.

Prediction: The estimate of an outcome obtained by plugging specific values of the explanatory variables into an estimated model, usually a multiple regression model.

Prediction Error: The difference between the actual outcome and a prediction of that outcome.

Prediction Interval: A confidence interval for an unknown outcome on a dependent variable in a multiple regression model.

Predictor Variable: *See* explanatory variable.

Probability Density Function (pdf): A function that, for discrete random variables, gives the probability that the random variable takes on each value; for continuous random variables, the area under the pdf gives the probability of various events.

Probability Limit: The value to which an estimator converges as the sample size grows without bound.

Probit Model: A model for binary responses where the response probability is the standard normal cdf evaluated at a linear function of the explanatory variables.

Program Evaluation: An analysis of a particular private or public program using econometric methods to obtain the causal effect of the program.

Proportionate Change: The change in a variable relative to its initial value; mathematically, the change divided by the initial value.

Proxy Variable: An observed variable that is related but not identical to an unobserved explanatory variable in multiple regression analysis.

Pseudo *R*-Squared: Any number of goodness-of-fit measures for limited dependent variable models.

Q

Quadratic Form: A mathematical function where the vector argument both pre- and post-multiplies a square, symmetric matrix.

Quadratic Functions: Functions that contain squares of one or more explanatory variables; they capture diminishing or increasing effects on the dependent variable.

Quasi-Demeaned Data: In random effects estimation for panel data, it is the original data in each time period minus a fraction of the time average; these calculations are done for each cross-sectional observation.

Quasi-Differenced Data: In estimating a regression model with AR(1) serial correlation, it is the difference between the current time period and a multiple of the previous time period, where the multiple is the parameter in the AR(1) model.

Quasi-Experiment: *See* natural experiment.

Quasi-Likelihood Ratio Statistic: A modification of the likelihood ratio statistic that accounts for possible distributional misspecification, as in a Poisson regression model.

Quasi-Maximum Likelihood Estimation (QMLE): Maximum likelihood estimation where the log-likelihood function may not correspond to the actual conditional distribution of the dependent variable.

R

Random Assignment: The process by which observations are assigned to the treatment and control groups completely at random (i.e. not as a function of any observable characteristics)

Randomized Controlled Trial (RCT): An experimental design in which a treatment group (given a policy) and a control group (no policy) are randomly selected from the population. Assuming no pre-treatment differences between these groups, any observed differences should be a result of the policy given to the treatment group.

Regression adjustment: A method for dealing with non-random assignment that involves additional control variables. The inclusion of these variables allows for identification of the causal effect of a policy.

***R*-Squared:** In a multiple regression model, the proportion of the total sample variation in the dependent variable that is explained by the independent variable.

***R*-Squared Form of the *F* Statistic:** The *F* statistic for testing exclusion restrictions expressed in terms of the *R*-squareds from the restricted and unrestricted models.

Random Coefficient (Slope) Model: A multiple regression model where the slope parameters are allowed to depend on unobserved unit-specific variables.

Random Effects Estimator: A feasible GLS estimator in the unobserved effects model where the unobserved effect is assumed to be uncorrelated with the explanatory variables in each time period.

Random Effects Model: The unobserved effects panel data model where the unobserved effect is assumed to be uncorrelated with the explanatory variables in each time period.

Random Sample: A sample obtained by sampling randomly from the specified population.

Random Sampling: A sampling scheme whereby each observation is drawn at random from the population. In particular, no unit is more likely to be selected than any other unit, and each draw is independent of all other draws.

Random Variable: A variable whose outcome is uncertain.

Random Vector: A vector consisting of random variables.

Random Walk: A time series process where next period's value is obtained as this period's value, plus an independent (or at least an uncorrelated) error term.

Random Walk with Drift: A random walk that has a constant (or drift) added in each period.

Rank Condition: A sufficient condition for identification of a model with one or more endogenous explanatory variables.

Rank of a Matrix: The number of linearly independent columns in a matrix.

Rational Distributed Lag (RDL) Model: A type of infinite distributed lag model where the lag distribution depends on relatively few parameters.

Reduced Form Equation: A linear equation where an endogenous variable is a function of exogenous variables and unobserved errors.

Reduced Form Error: The error term appearing in a reduced form equation.

Reduced Form Parameters: The parameters appearing in a reduced form equation.

Regressand: *See* dependent variable.

Regression Specification Error Test (RESET): A general test for functional form in a multiple regression model; it is an F test of joint significance of the squares, cubes, and perhaps higher powers of the fitted values from the initial OLS estimation.

Regression through the Origin: Regression analysis where the intercept is set to zero; the slopes are obtained by minimizing the sum of squared residuals, as usual.

Regressor: *See* explanatory variable.

Rejection Region: The set of values of a test statistic that leads to rejecting the null hypothesis.

Rejection Rule: In hypothesis testing, the rule that determines when the null hypothesis is rejected in favor of the alternative hypothesis.

Relative Change: *See* proportionate change.

Resampling Method: A technique for approximating standard errors (and distributions of test statistics) whereby a series of samples are obtained from the original data set and estimates are computed for each subsample.

Residual: The difference between the actual value and the fitted (or predicted) value; there is a residual for each observation in the sample used to obtain an OLS regression line.

Residual Analysis: A type of analysis that studies the sign and size of residuals for particular observations after a multiple regression model has been estimated.

Residual Sum of Squares: *See* sum of squared residuals.

Response Probability: In a binary response model, the probability that the dependent variable takes on the value one, conditional on explanatory variables.

Response Variable: *See* dependent variable.

Restricted Model: In hypothesis testing, the model obtained after imposing all of the restrictions required under the null.

Retrospective Data: Data collected based on past, rather than current, information.

Root Mean Squared Error (RMSE): Another name for the standard error of the regression in multiple regression analysis.

Row Vector: A vector of numbers arranged as a row.

S

Sample Average: The sum of n numbers divided by n; a measure of central tendency.

Sample Correlation Coefficient: An estimate of the (population) correlation coefficient from a sample of data.

Sample Covariance: An unbiased estimator of the population covariance between two random variables.

Sample Regression Function (SRF): *See* OLS regression line.

Sample Standard Deviation: A consistent estimator of the population standard deviation.

Sample Variance: An unbiased, consistent estimator of the population variance.

Sampling Distribution: The probability distribution of an estimator over all possible sample outcomes.

Sampling Standard Deviation: The standard deviation of an estimator, that is, the standard deviation of a sampling distribution.

Sampling Variance: The variance in the sampling distribution of an estimator; it measures the spread in the sampling distribution.

Scalar Multiplication: The algorithm for multiplying a scalar (number) by a vector or matrix.

Scalar Variance-Covariance Matrix: A variance-covariance matrix where all off-diagonal terms are zero and the diagonal terms are the same positive constant.

Score Statistic: *See* Lagrange multiplier statistic.

Seasonal Dummy Variables: A set of dummy variables used to denote the quarters or months of the year.

Seasonality: A feature of monthly or quarterly time series where the average value differs systematically by season of the year.

Seasonally Adjusted: Monthly or quarterly time series data where some statistical procedure—possibly regression on seasonal dummy variables—has been used to remove the seasonal component.

Selected Sample: A sample of data obtained not by random sampling but by selecting on the basis of some observed or unobserved characteristic.

Self-Selection: Deciding on an action based on the likely benefits, or costs, of taking that action.

Self-Selection Problem: Occurs when there is non-random assignment and inclusion in the treatment and control group systematically depends on individual characteristics.

Semi-Elasticity: The percentage change in the dependent variable given a one-unit increase in an independent variable.

Sensitivity Analysis: The process of checking whether the estimated effects and statistical significance of key explanatory variables are sensitive to inclusion of other explanatory variables, functional form, dropping of potentially outlying observations, or different methods of estimation.

Sequentially Exogenous: A feature of an explanatory variable in time series (or panel data) models where the error term in the current time period has a zero mean conditional on all current and past explanatory variables; a weaker version is stated in terms of zero correlations.

Serial Correlation: In a time series or panel data model, correlation between the errors in different time periods.

Serial Correlation-Robust Standard Error: A standard error for an estimator that is (asymptotically) valid whether or not the errors in the model are serially correlated.

Serially Uncorrelated: The errors in a time series or panel data model are pairwise uncorrelated across time.

Short-Run Elasticity: The impact propensity in a distributed lag model when the dependent and independent variables are in logarithmic form.

Significance Level: The probability of a Type I error in hypothesis testing.

Simple Linear Regression Model: A model where the dependent variable is a linear function of a single independent variable, plus an error term.

Simultaneity: A term that means at least one explanatory variable in a multiple linear regression model is determined jointly with the dependent variable.

Simultaneity Bias: The bias that arises from using OLS to estimate an equation in a simultaneous equations model.

Simultaneous Equations Model (SEM): A model that jointly determines two or more endogenous variables, where each endogenous variable can be a function of other endogenous variables as well as of exogenous variables and an error term.

Skewness: A measure of how far a distribution is from being symmetric, based on the third moment of the standardized random variable.

Slope Parameter: The coefficient on an independent variable in a multiple regression model.

Smearing Estimate: A retransformation method particularly useful for predicting the level of a response variable when a linear model has been estimated for the natural log of the response variable.

Spreadsheet: Computer software used for entering and manipulating data.

Spurious Regression Problem: A problem that arises when regression analysis indicates a relationship between two or more unrelated time series processes simply because each has a trend, is an integrated time series (such as a random walk), or both.

Stable AR(1) Process: An AR(1) process where the parameter on the lag is less than one in absolute value. The correlation between two random variables in the sequence declines to zero at a geometric rate as the distance between the random variables increases, and so a stable AR(1) process is weakly dependent.

Standard Deviation: A common measure of spread in the distribution of a random variable.

Standard Deviation of $\hat{\beta}_j$: A common measure of spread in the sampling distribution of $\hat{\beta}_j$.

Standard Error of $\hat{\beta}_1$: The standard error of the OLS slope estimator. In the simple regression model, this is $SE(\hat{\beta}_1) = \hat{\sigma}/\sqrt{SST_x}$.

Standard Error of $\hat{\beta}_j$: An estimate of the standard deviation in the sampling distribution of $\hat{\beta}_j$.

Standard Error of the Regression (SER): In multiple regression analysis, the estimate of the standard deviation of the population error, obtained as the square root of the sum of squared residuals over the degrees of freedom.

Standardized Coefficients: Regression coefficients that measure the standard deviation change in the dependent variable given a one standard deviation increase in an independent variable.

Standardized Random Variable: A random variable transformed by subtracting off its expected value and dividing the result by its standard deviation; the new random variable has mean zero and standard deviation one.

Static Model: A time series model where only contemporaneous explanatory variables affect the dependent variable.

Stationary Process: A time series process where the marginal and all joint distributions are invariant across time.

Statistical Significance: The importance of an estimate as measured by the size of a test statistic, usually a t statistic.

Statistically Insignificant: Failure to reject the null hypothesis that a population parameter is equal to zero, at the chosen significance level.

Statistically Significant: Rejecting the null hypothesis that a parameter is equal to zero against the specified alternative, at the chosen significance level.

Stochastic Process: A sequence of random variables indexed by time.

Stratified Sampling: A nonrandom sampling scheme whereby the population is first divided into several nonoverlapping, exhaustive strata, and then random samples are taken from within each stratum.

Strict Exogeneity: An assumption that holds in a time series or panel data model when the explanatory variables are strictly exogenous.

Strictly Exogenous: A feature of explanatory variables in a time series or panel data model where the error term at any time period has zero expectation, conditional on the explanatory variables in all time periods; a less restrictive version is stated in terms of zero correlations.

Strongly Dependent: *See* highly persistent.

Structural Equation: An equation derived from economic theory or from less formal economic reasoning.

Structural Error: The error term in a structural equation, which could be one equation in a simultaneous equations model.

Structural Parameters: The parameters appearing in a structural equation.

Studentized Residuals: The residuals computed by excluding each observation, in turn, from the estimation, divided by the estimated standard deviation of the error.

Sum of Squared Residuals (SSR): In multiple regression analysis, the sum of the squared OLS residuals across all observations.

Summation Operator: A notation, denoted by $\sum$, used to define the summing of a set of numbers.

Symmetric Distribution: A probability distribution characterized by a probability density function that is symmetric around its median value, which must also be the mean value (whenever the mean exists).

Symmetric Matrix: A (square) matrix that equals its transpose.

T

***t* Distribution:** The distribution of the ratio of a standard normal random variable and the square root of an independent chi-square random variable, where the chi-square random variable is first divided by its *df*.

***t* Ratio:** *See t* statistic.

***t* Statistic:** The statistic used to test a single hypothesis about the parameters in an econometric model.

Test Statistic: A rule used for testing hypotheses where each sample outcome produces a numerical value.

Text Editor: Computer software that can be used to edit text files.

Text (ASCII) File: A universal file format that can be transported across numerous computer platforms.

Time-Demeaned Data: Panel data where, for each cross-sectional unit, the average over time is subtracted from the data in each time period.

Time Series Data: Data collected over time on one or more variables.

Time Series Process: *See* stochastic process.

Time Trend: A function of time that is the expected value of a trending time series process.

Tobit Model: A model for a dependent variable that takes on the value zero with positive probability but is roughly continuously distributed over strictly positive values. (*See also* corner solution response.)

Top Coding: A form of data censoring where the value of a variable is not reported when it is above a given threshold; we only know that it is at least as large as the threshold.

Total Sum of Squares (SST): The total sample variation in a dependent variable about its sample average.

Trace of a Matrix: For a square matrix, the sum of its diagonal elements.

Transpose: For any matrix, the new matrix obtained by interchanging its rows and columns.

Treatment Group: In program evaluation, the group that participates in the program.

Trend-Stationary Process: A process that is stationary once a time trend has been removed; it is usually implicit that the detrended series is weakly dependent.

True Model: The actual population model relating the dependent variable to the relevant independent variables, plus a disturbance, where the zero conditional mean assumption holds.

Truncated Normal Regression Model: The special case of the truncated regression model where the underlying population model satisfies the classical linear model assumptions.

Truncated Regression Model: A linear regression model for cross-sectional data in which the sampling scheme entirely excludes, on the basis of outcomes on the dependent variable, part of the population.

Truncation Lag: A parameter in HAC standard errors that determines the number of lags of the residuals that need to be included to correct for serial correlation.

Two-Sided Alternative: An alternative where the population parameter can be either less than or greater than the value stated under the null hypothesis.

Two Stage Least Squares (2SLS) Estimator: An instrumental variables estimator where the IV for an endogenous explanatory variable is obtained as the fitted value from regressing the endogenous explanatory variable on all exogenous variables.

Two-Tailed Test: A test against a two-sided alternative.

Type I Error: A rejection of the null hypothesis when it is true.

Type II Error: The failure to reject the null hypothesis when it is false.

U

Unbalanced Panel: A panel data set where certain years (or periods) of data are missing for some cross-sectional units.

Unbiased Estimator: An estimator whose expected value (or mean of its sampling distribution) equals the population value (regardless of the population value).

Uncentered *R*-squared: The *R*-squared computed without subtracting the sample average of the dependent variable when obtaining the total sum of squares (SST).

Unconditional Forecast: A forecast that does not rely on knowing, or assuming values for, future explanatory variables.

Uncorrelated Random Variables: Random variables that are not linearly related.

Underspecifying a Model: *See* excluding a relevant variable.

Unidentified Equation: An equation with one or more endogenous explanatory variables where sufficient instrumental variables do not exist to identify the parameters.

Unconfounded Assignment: *See* conditional independence.

Unit Root Process: A highly persistent time series process where the current value equals last period's value, plus a weakly dependent disturbance.

Unobserved Effect: In a panel data model, an unobserved variable in the error term that does not change over time. For cluster samples, an unobserved variable that is common to all units in the cluster.

Unit Roots: In the time series process $y_t = \alpha + \rho y_{t-1} + u_t$, we say that y_t has a unit root if $\rho = 1$. This is also known as a random walk and is an unpredictable process.

Unobserved Effects Model: A model for panel data or cluster samples where the error term contains an unobserved effect.

Unobserved Heterogeneity: *See* unobserved effect.

Unrestricted Model: In hypothesis testing, the model that has no restrictions placed on its parameters.

Upward Bias: The expected value of an estimator is greater than the population parameter value.

V

Variance-Covariance Matrix: For a random vector, the positive semi-definite matrix defined by putting the variances down the diagonal and the covariances in the appropriate off-diagonal entries.

Variance-Covariance Matrix of the OLS Estimator: The matrix of sampling variances and covariances for the vector of OLS coefficients.

Variance Inflation Factor: In multiple regression analysis under the Gauss-Markov assumptions, the term in the sampling variance affected by correlation among the explanatory variables.

Variance of the Prediction Error: The variance in the error that arises when predicting a future value of the dependent variable based on an estimated multiple regression equation.

Vector Autoregressive (VAR) Model: A model for two or more time series where each variable is modeled as a linear function of past values of all variables, plus disturbances that have zero means given all past values of the observed variables.

W

Wald Statistic: A general test statistic for testing hypotheses in a variety of econometric settings; typically, the Wald statistic has an asymptotic chi-square distribution.

Weak Instruments: Instrumental variables that are only slightly correlated with the relevant endogenous explanatory variable or variables.

Weakly Dependent: A term that describes a time series process where some measure of dependence between random variables at two points in time—such as correlation—diminishes as the interval between the two points in time increases.

Weighted Least Squares: Refer to below.

Weighted Least Squares (WLS) Estimator: An estimator used to adjust for a known form of heteroskedasticity, where each squared residual is weighted by the inverse of the (estimated) variance of the error.

White Test for Heteroskedasticity: A test for heteroskedasticity in which the squared residuals are regressed on linear and non-linear functions of the explanatory variables.

Within Estimator: *See* fixed effects estimator.

Within Transformation: *See* fixed effects transformation.

Y

Year Dummy Variables: For data sets with a time series component, dummy (binary) variables equal to one in the relevant year and zero in all other years.

Z

Zero Conditional Mean Assumption: A key assumption used in multiple regression analysis that states that, given any values of the explanatory variables, the expected value of the error equals zero. (*See* Assumptions MLR.4, TS.3, and TS.3' in the text.)

Zero Matrix: A matrix where all entries are zero.

Zero-One Variable: *See* dummy variable.

Index

Numbers

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z