

Programming with Data

Session 1: Introduction

Dr. Wang Jiwei

Master of Professional Accounting

About the course

What will this course cover?



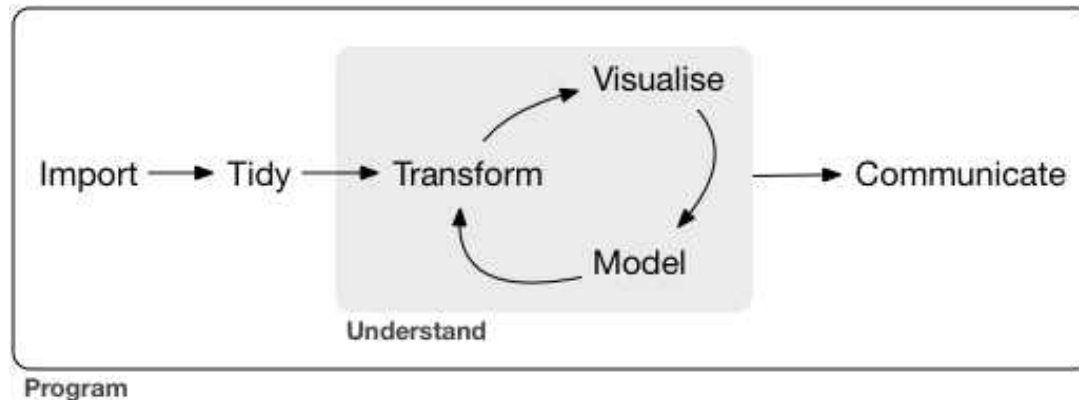
1. Prerequisite
 - University Statistics
2. Programming with R
 - R programming foundations
3. Linear regressions with R
 - Forecast/Predict financial outcomes
4. Binary classification with R
 - Event prediction
 - Classification/detection (of financial fraud)
5. Data visualizations with R
 - ggplot2 package in R
6. Advanced methods
 - Lasso, Ridge and Elastic Net regressions
 - Introduction to machine learning

| Using R for forecasting and forensics

Teaching philosophy

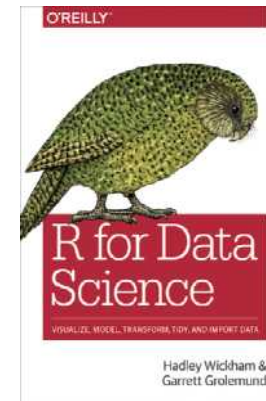
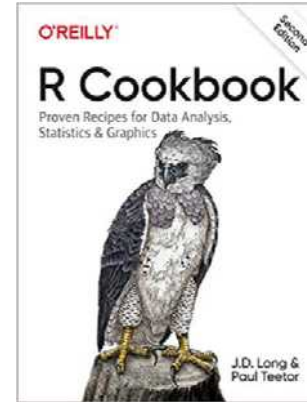
1. Programming is best learned by doing it
 - more thinking and hands-on practising
2. Working with others greatly extends learning
 - If you are ahead:
 - The best sign that you've mastered a topic is if you can explain it to others
 - If you are lost:
 - Gives you a chance to get the help you need
3. We generally follow the following model to learn programming with data

Source: R for Data Science



Textbook and learning materials

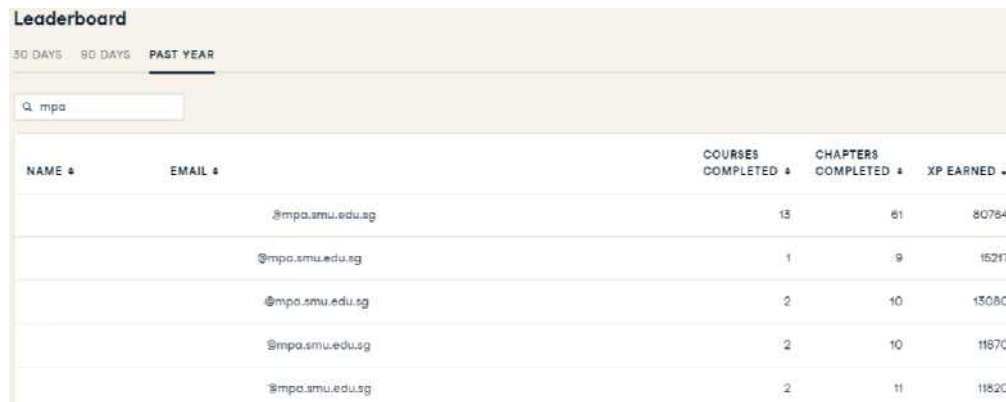
- All course materials on SMU eLearn
- There is no required textbook
- If you prefer having a textbook...
 - **R Cookbook** is good for beginners
 - **R for Data Science** is good for more advanced learners
- Announcements will be mainly on eLearn
- Other useful websites
 - <https://www.r-bloggers.com/>
 - <https://stackoverflow.com/questions>
 - <https://www.google.com/>



Self learning and Datacamp

- You are encouraged to go beyond the assigned materials, either through Datacamp or other online learning platforms such as Coursera and Udemy.
- Datacamp is providing *free* access to their *full* library of analytics and coding online tutorials
 - You will have free access for 6 months (July 1 to Dec 30, 2021), subject to renewal
- Suggestion: enroll into the **"Data Analysts with R/Python"** career track on Datacamp and finish all courses before completing your degree
 - Check eLearn for link to access Datacamp for free
 - Datacamp automatically records when you finish these

Practice! Practice! Practice!



		COURSES COMPLETED	CHAPTERS COMPLETED	XP EARNED
NAME	EMAIL			
	@mpa.smu.edu.sg	13	61	80764
	@mpa.smu.edu.sg	1	9	15217
	@mpa.smu.edu.sg	2	10	13080
	@mpa.smu.edu.sg	2	10	11970
	@mpa.smu.edu.sg	2	11	11820

Grading

- Participation @ 20%
- Progress assessment @ 30%
- Group project @ 50%
- There is no final exam

Must attempt all components and must pass all components to pass the course

source: medium.com



Participation

In Class

- Come to class to earn 50%
 - If you have a conflict, email me
 - Excused classes do not impact your participation grade
- Ask questions to **extend** or **clarify**
- Answer questions and explain answers
 - Give it your best shot!
- Help those in your group to understand concepts
- Present your work to the class
- Always **on your camera** and speak with your microphone
- Other initiatives to enrich the classroom learning experience

Outside of Class

- Verify your understanding of the material
- Apply to other real world data
 - Techniques and code will be useful after graduation
- Answers to assignments are expected to be your own work, unless otherwise stated
 - No sharing answers (unless otherwise stated)
- All submissions on eLearn
 - on time and follow instructions
- I will provide snippets of code to help you with trickier parts

Group project

- Data science competition format, hosted on **Kaggle** or similar platforms.
- The project will finish in Session 10 with group presentations
- I will give your more details in a separate document

kaggle

Expectations

In class:

- Participate
 - Ask questions
 - Clarify
 - Add to the discussion
 - Answer questions
 - Work with classmates



Outside of class:

- Check eLearn for course announcements
- Do the tutorials on Datacamp if you are not familiar with R
 - This will make the course much easier!
- Do individual work on your own (unless otherwise stated)
 - Submit on eLearn
- Do online courses through Datacamp or other platforms
- Office hours are there to help!
 - Short questions can be emailed instead

Office hours

- Appointment at the following link
 - <https://calendly.com/drdataking>
 - The default time is 15 minutes
 - If you want longer time, you may book multiple slots
- Short questions can be emailed
 - I try to respond within 24 hours
- Teaching Assistant (check eLearn)
 - always make appointment before approaching TA



Tech use

- Laptops and other tech are OK!
 - Use them for learning and course related
- Examples of good tech use:
 - Taking notes
 - Viewing slides
 - Working out problems
 - Group discussion
- Avoid:
 - Messaging your friends on Whatsapp/Wechat/Telegram/etc
 - Working on homework or group project in class
 - Playing games or watching livestreams
- **In-class cellphone and laptop use lowers exam scores**



About you

Introduction to analytics

What is analytics?

Oxford: a careful and complete analysis of data using a model, usually performed by a computer; information resulting from this analysis

Webster: the method of logical analysis

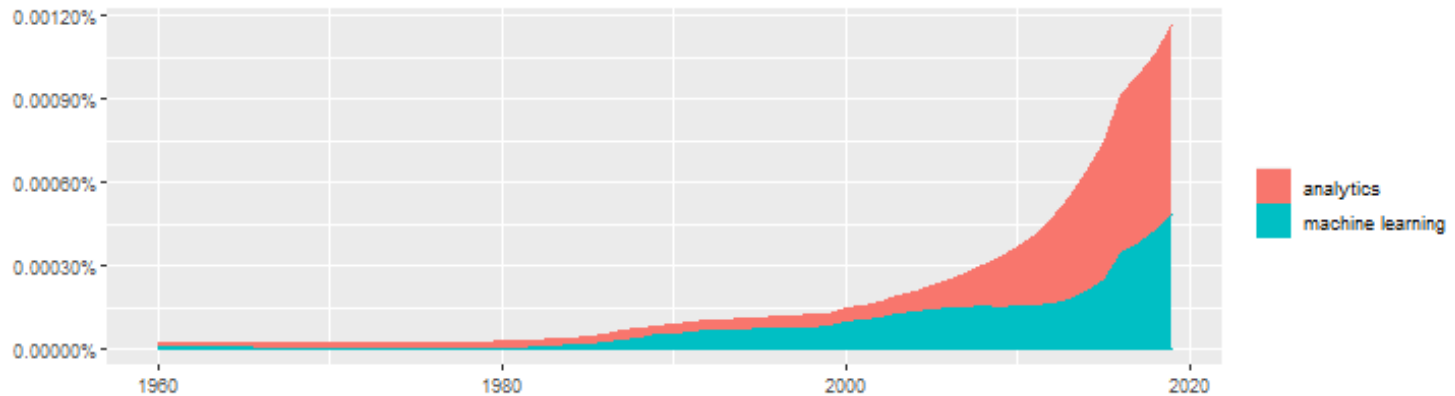
Wikipedia: the discovery, interpretation, and communication of meaningful patterns in data and applying those patterns towards effective decision making

Simply put: Solving problems using data

- Additional layers we can add to the definition:
 - Solving problems using *a lot of* data
 - Solving problems using data *and statistics*
 - Solving problems using data *and computers (programming and/or specialized software)*

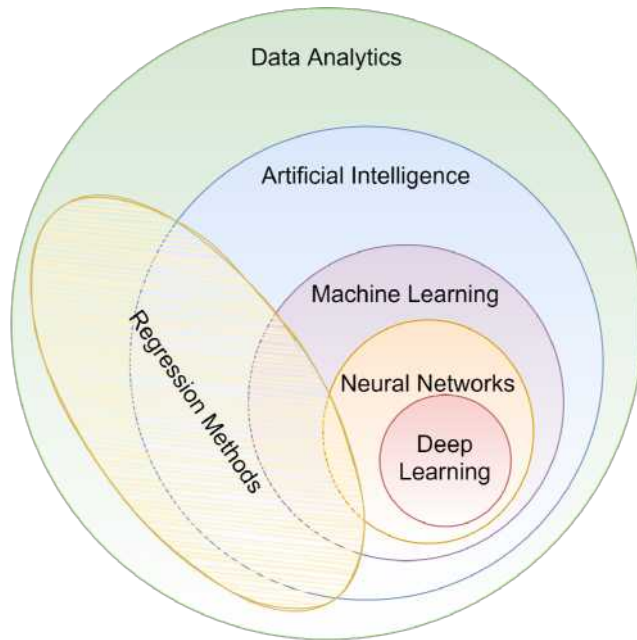
The trend

We search "analytics" in Google Books and display the graph showing how the word has occurred since 1960



Made using R `package:seancarmody/ngramr` which is available on CRAN (the central depository for R packages) and can be installed from RStudio directly

Analytics vs AI/machine learning



- In class reading:
 - Future of everything: **AI Will Enhance Us, Not Replace Us**
 - "The future isn't AI versus humans. It is AI-enhanced humans doing what humans are best at."
 - AI Ethics: **Apple Card is facing a formal investigation**
 - "We need transparency and fairness."

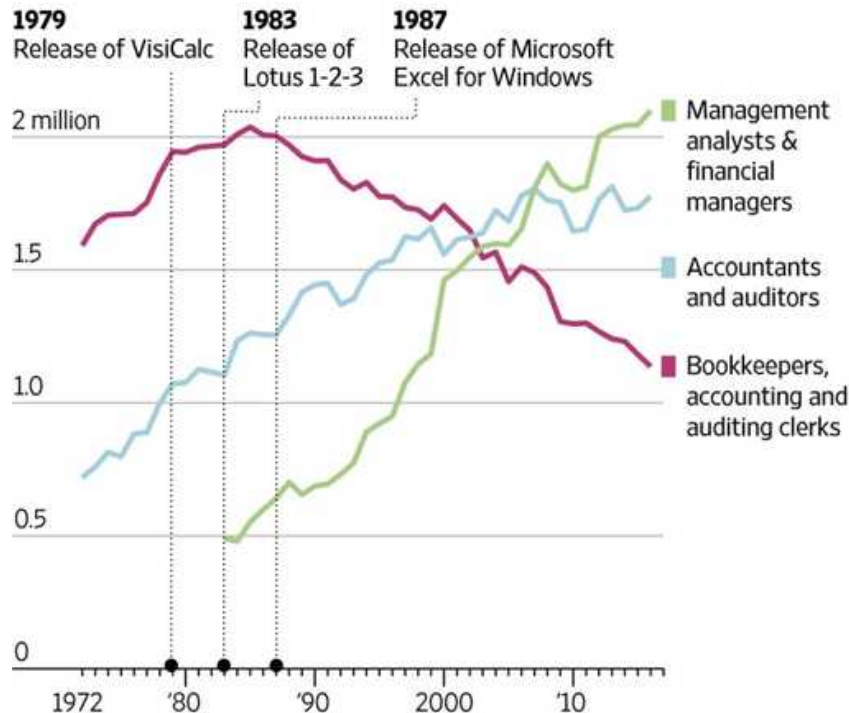
■ Class Discussion

How will Analytics/AI/ML change society and the accounting profession?

What happened before?

The Spreadsheet Apocalypse, Revisited

Jobs in bookkeeping plummeted after the introduction of spreadsheet software, but jobs in accounting and analysis took off.



Notes: There is no data for 1982. Changes in occupational definitions in 1983, 2000 and 2011 mean that data is not strictly comparable across time. There was no category for management analysts or financial managers prior to 1983.

Source: Bureau of Labor Statistics

THE WALL STREET JOURNAL.

Who uses analytics?

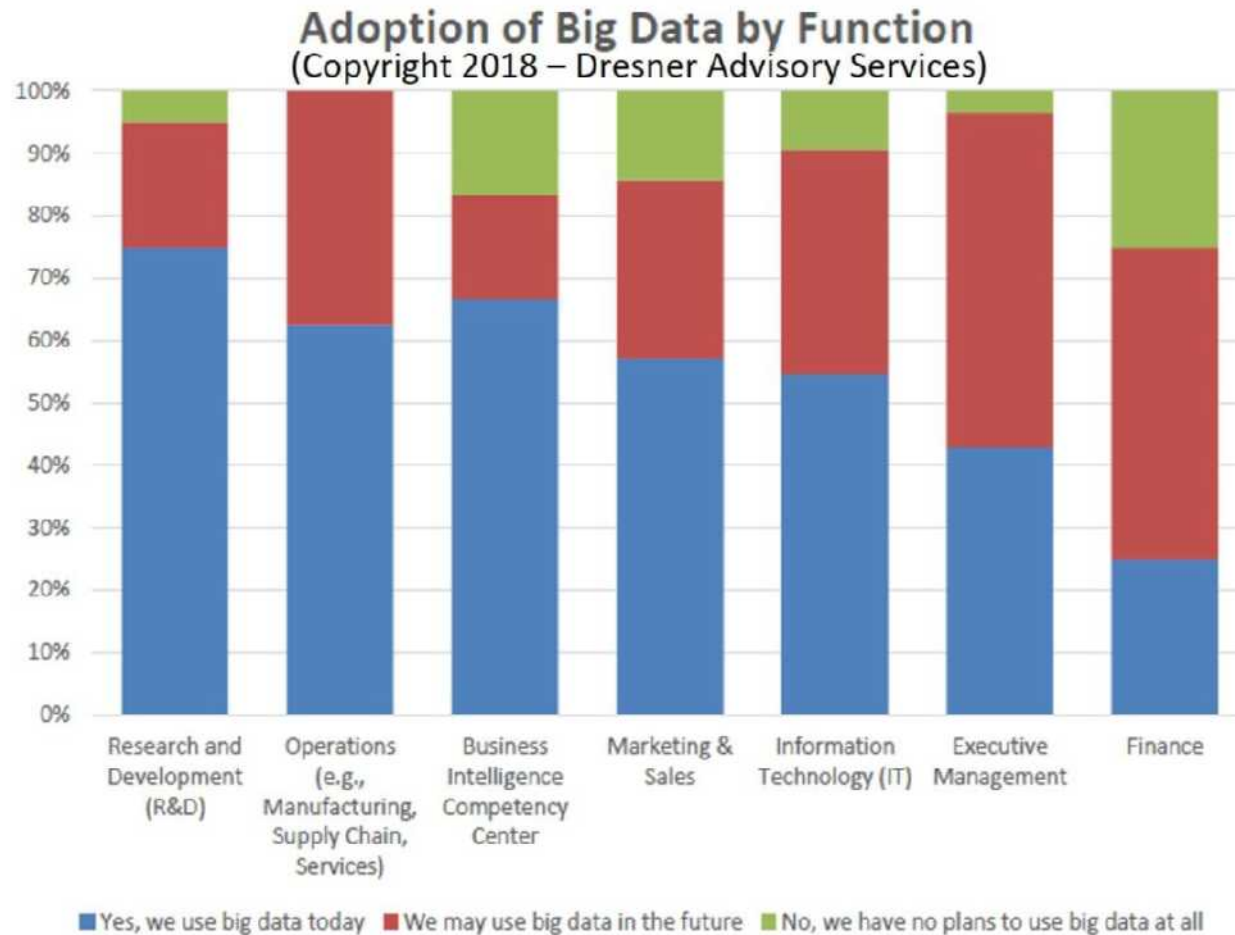
In general

- Companies
 - Finance
 - Manufacturing
 - Transportation
 - Computing
 - ...
- Governments
 - AI.Singapore
 - Big data office
 - "Smart" initiatives
- Academics
- Individuals!

59% of companies where using big data in a 2018 survey!

Which corporate function has the highest/lowest adoption of big data analytics?

Adoption of big data by function



What analytics for?

- Customer service
 - Royal Bank of Scotland
 - Understanding customer complaints
- Improving products
 - Siemens' Internet of Trains
 - Improving train reliability
- Auditing
 - Continuous Auditing at DBS
 - The Future of Auditing is Auditing the Future
- How about your company?



SIEMENS



State of business analytics?

- **Dresner Advisory Service's 2018 Market Study**
- Executive Management, Operations, Sales and Finance are the four primary roles driving business analytics adoption in 2018.
- Dashboards, reporting, end-user self-service, advanced visualization, and data warehousing are the top five most important initiatives.

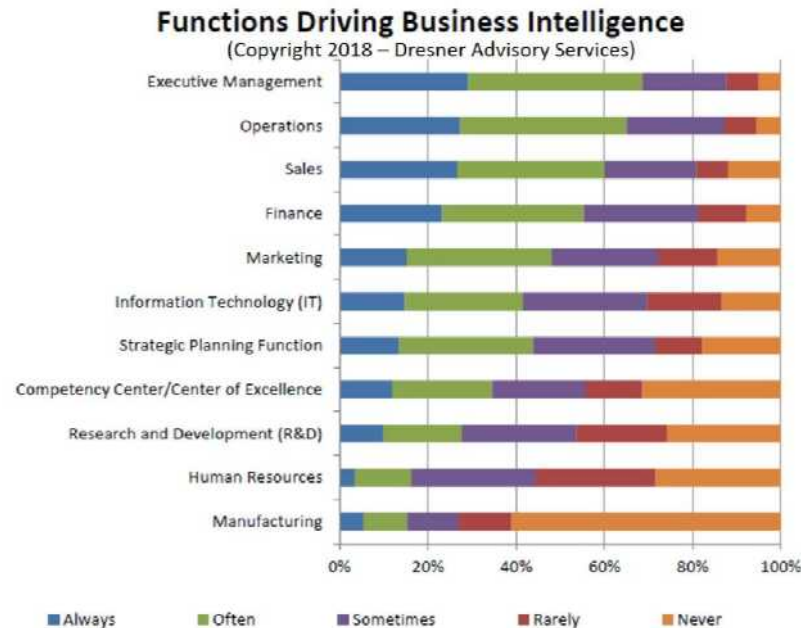


Figure 5 – Functions driving business intelligence

Head of Finance Data

Key tasks and responsibilities

- Leading the Finance Data Team to maintain and improve the Financial data application landscape (BI Reporting, Planning and Budgeting system) and data pipelines powering Finance systems and reporting.
- Enable business users to further improve the data literacy and ultimately drive data decision making.



Qualifications & Skills

- Degree in Accounting, Finance, Business Administration, Computer Science or related field
- Experience in big database including strong expertise in SQL
- Excel, R/Python (plus), SAP hands-on experience
- Creative and analytical thinker with strong problem-solving skills
- Strong written and oral communication skills

Statistics Foundations

Frequentist vs Bayesian statistics

Frequentist statistics

A specific test is one of an infinite number of replications

- The "correct" answer should occur most frequently, i.e., with a high probability
- Focus on true vs. false
- Treat unknowns as fixed constants to figure out
 - Not random quantities
- Where it's used
 - Classical statistics methods
 - Like OLS

Bayesian statistics

Focus on distributions and beliefs

- Prior distribution -- what is believed before the experiment
- Posterior distribution: an updated belief of the distribution due to the experiment
- Derive distributions of parameters
- Where it's used:
 - Many machine learning methods
 - Bayesian updating acts as the learning
 - Bayesian statistics

Frequentist: Repeat the test

| Did the sun explode just now?

```
# Don't worry, we will learn how to program in R soon.  
# Define a detector  
# repeat the test with frequentist statistics  
  
detector <- function() {  
  dice <- sample(1:6, size = 2, replace = TRUE)  
  if (sum(dice) == 12) {  
    "exploded"  
  } else {  
    "still there"  
  }  
}  
  
experiment <- replicate(1000, detector())  
# p value  
paste("p-value: ",  
      sum(experiment == "still there") / 1000,  
      "-- Failed to reject H_0 that sun didn't explode")
```

```
## [1] "p-value: 0.968 -- Failed to reject H_0 that sun didn't explode"
```

| Frequentist: The sun didn't explode

Bayesian: Bayes rule

Did the sun explode just now?

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- A : The sun exploded
- B : The detector said it exploded
- $P(A)$: Really, really small. Say, ~ 0 . Prior belief
- $P(B)$: $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$. Experiment
- $P(B|A)$: $\frac{35}{36}$. Post belief

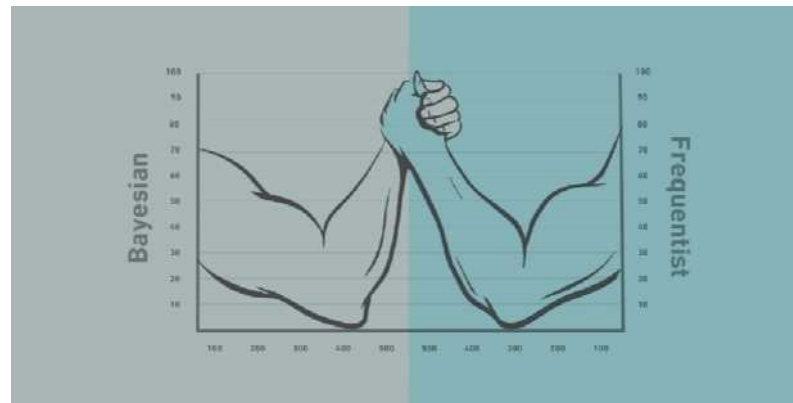
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\frac{35}{36} \times \sim 0}{\frac{1}{36}} = 35 \times \sim 0 \approx 0$$

Bayesian: The sun didn't explode

What analytics typically relies on

- Regression approaches
 - Most often done in a frequentist manner
 - Can be done in a Bayesian manner as well
- Machine learning
 - Sometimes Bayesian, sometime frequentist

We will mainly use frequentist statistics and some applications in bayesian -- for our purposes, we will not debate the merits of either school of thought, but use tools derived from both



Confusion from frequentist approaches

- Possible contradictions:
 - F test says the model is good yet nothing is statistically significant
 - Individual p -values are good yet the model isn't
 - One measure says the model is good yet another doesn't

There are many ways to measure a model, each with their own merits. They don't always agree, and it's on us to pick a reasonable measure. We will discuss more in applications.

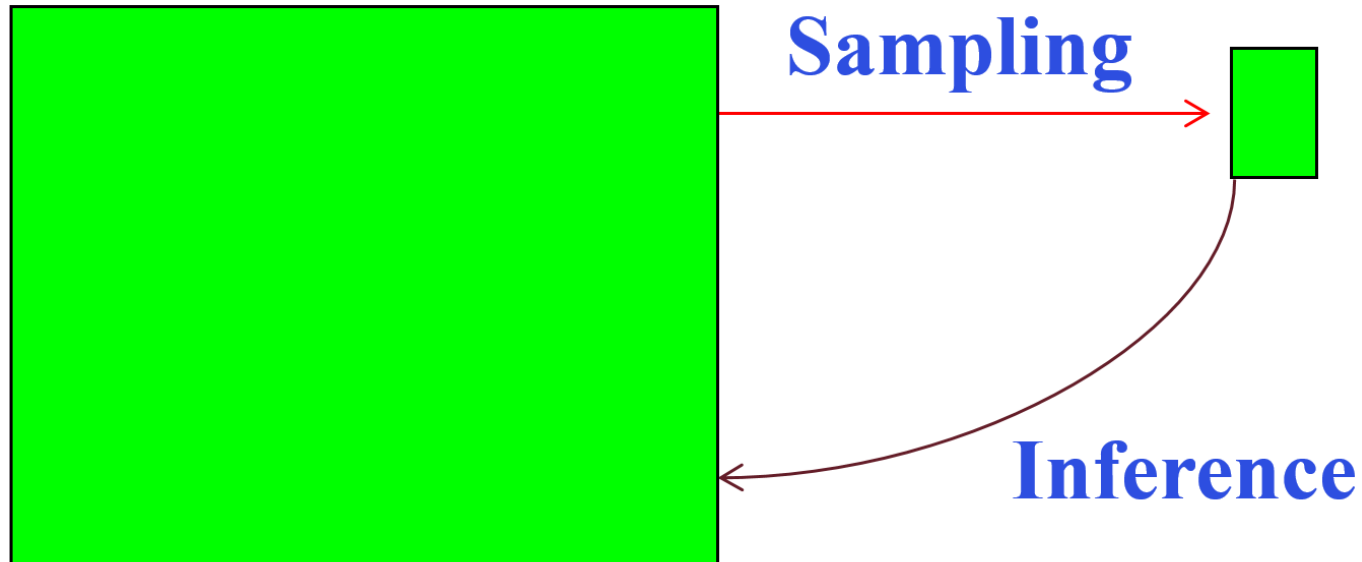
Frequentist approaches to things

Population vs Sample

- Population: all objects belonging to a specified set
 - e.g., All companies in Singapore
- Sample: a (random) subset of the population

Population

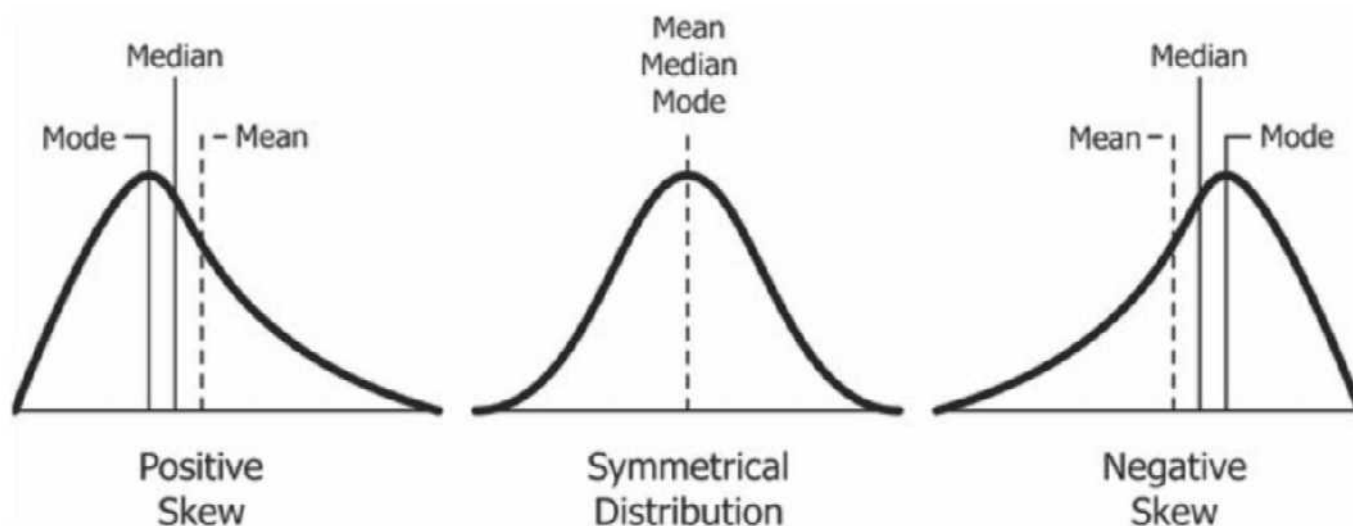
Sample



Parameters vs statistics

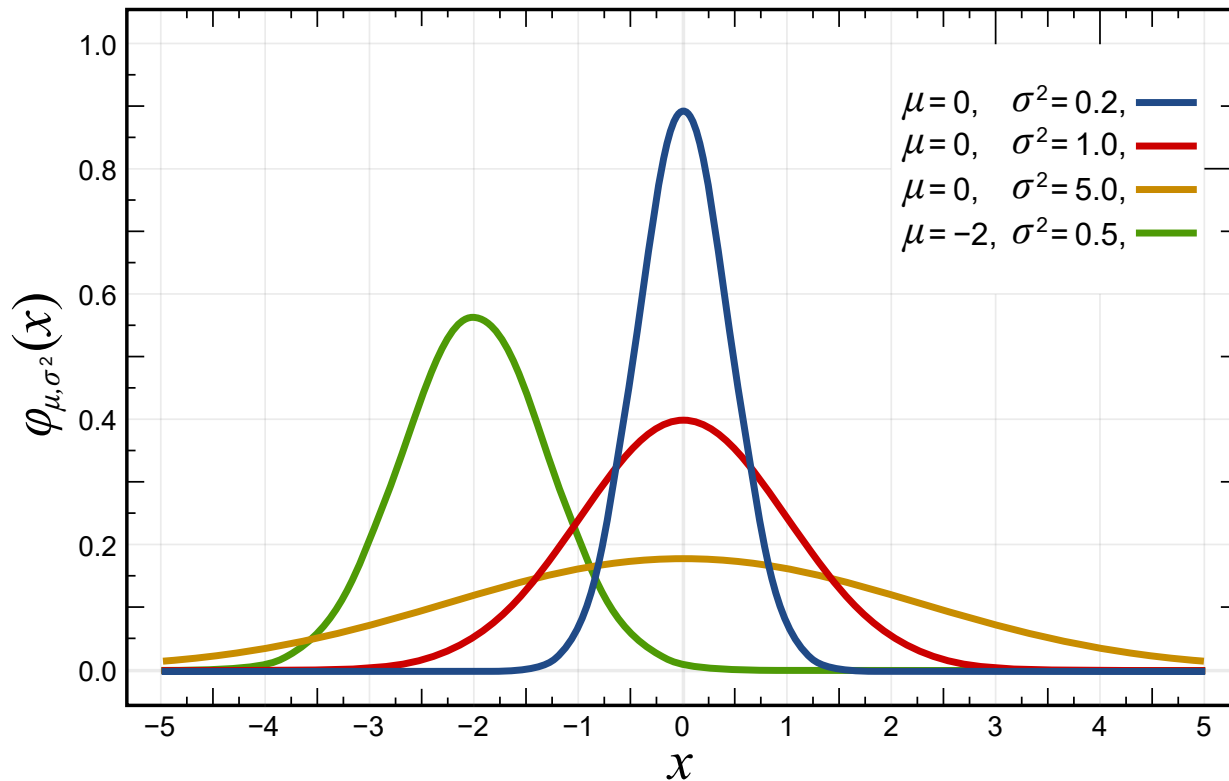
Population parameters vs Sample statistics of a given variable (such as height of boys or earning of companies)

- mean
- median/quantile
- mode
- standard deviation/variance
- max/min
- distribution



Normal distribution

A normal (gaussian or bell curve) distribution is a type of continuous probability distribution for a real-valued random variable with the same values of mean, median and mode.



Sampling error

- Sample statistic will *not* be exactly equal to population parameter
 - But should be close
- How close depends on sample size
 - Confidence interval

$$\bar{X} \pm Z_{1-\alpha/2} \frac{\sigma}{\sqrt{N}}$$

where \bar{X} is the sample mean, $Z_{1-\alpha/2}$ is the $1 - \alpha/2$ critical value of the standard normal distribution (1.68, 1.96 and 2.58 for 10%, 5%, and 1% respectively, which corresponding to confidence level of 90%, 95% and 99%), σ is the known population standard deviation, and N is the sample size.

- The larger the sample size N , the closer the sample statistic to the population parameter
 - trade-off between data collection costs and sample/margin error
 - we typically choose a confidence level (such as 99%) and a margin error to determine the minimum random sample size N

Law of large numbers

The law of large numbers states that as a sample size grows, its mean gets closer to the average of the whole population.

- Roll a 6 sided dice (1 to 6), the expected mean is 3.5

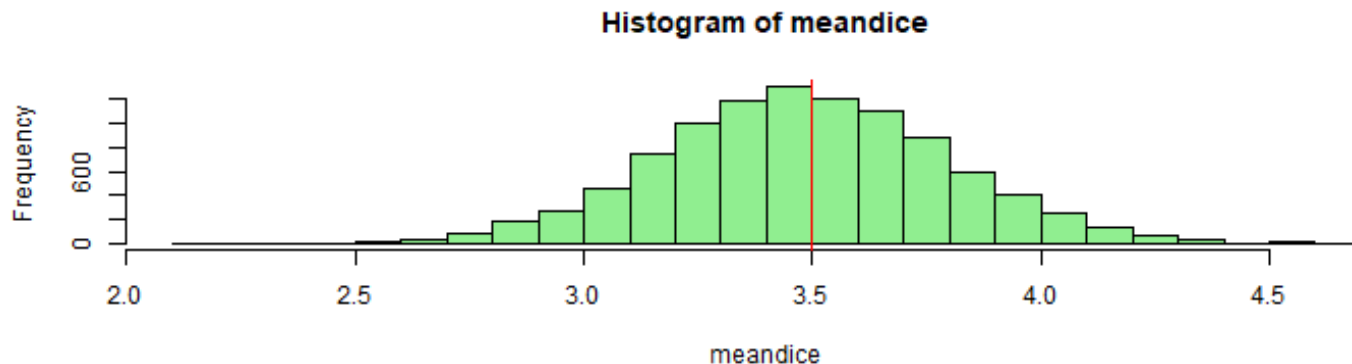
```
# Roll a dice
i <- 1
dice <- 0
times <- 10000
while (i <= times) {
  dice <- dice + sample(1:6, 1)
  i <- i + 1
}
paste("Roll", times, "times dice and the mean is", dice/times)
```

```
## [1] "Roll 10000 times dice and the mean is 3.5095"
```

Central Limit Theorem

If you take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

```
i <- 0; meandice <- c()
while (i <= 10000) {
  meandice <- append(meandice,
                    mean(sample(1:6, 30, replace = TRUE)))
  i <- i + 1
}
hist(meandice, col = "lightgreen", breaks = 20)
abline(v = 3.5, col = "blue")
abline(v = mean(meandice), col = "red")
```



Hypotheses

- H_0 : Null hypothesis
 - The status quo is correct
 - Your proposed model/prediction doesn't work
- H_A or H_1 : Alternative hypothesis
 - The model/prediction you are proposing works
- Frequentist statistics can never directly support H_0 !
 - Reject H_0 (a.k.a find Support for H_A) if $p\text{-value} < \text{a significance level}$ (such as 5% or 1%)
 - Fail to reject H_0 (a.k.a fail to find Support for H_A) if $p\text{-value} \geq \text{a significance level}$
- We can roughly understand $p\text{-value}$ as the probability of H_0
 - We will discuss more later

Even if our $p\text{-value}$ is 1, we can't say that the results prove the null hypothesis!

OLS terminology

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \varepsilon$$

$$\hat{y} = \alpha + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \dots + \hat{\varepsilon}$$

- y : The output in our model
 - dependent variable
 - predicted value
- \hat{y} : The *estimated* output in our model
- x_i : An input in our model
 - independent variables
 - features
 - predictors
- \hat{x}_i : An *estimated* input in our model
- $\hat{\cdot}$: Something *estimated*, "caret" or "hat"
- α : A constant, the expected value of y when all x_i are 0
- β_i : A coefficient on an input to our model
- ε : The error term
 - This is also the *residual* from the regression
 - What's left if you take actual y minus the model prediction

OLS statistical properties

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \varepsilon$$

$$\hat{y} = \alpha + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \dots + \hat{\varepsilon}$$

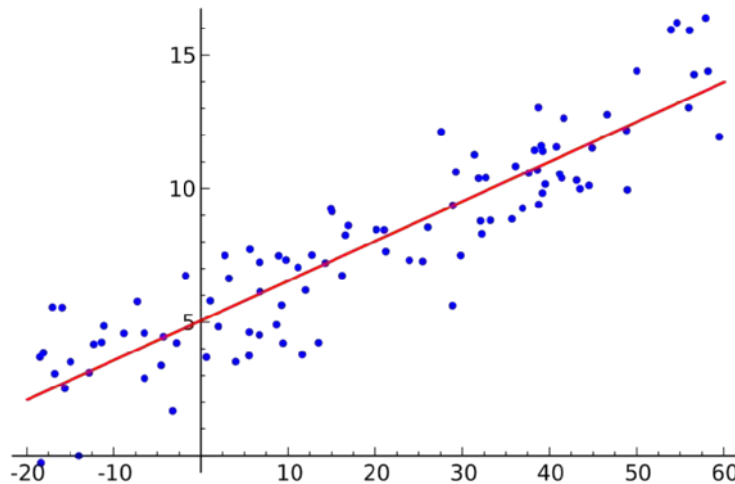
1. There should be a *linear* relationship between y and each x_i
 - i.e., y is [approximated by] a constant multiple of each x_i
 - Otherwise we **shouldn't** use a *linear* regression
2. Each \hat{x}_i is normally distributed
 - Not so important with larger data sets, but a good to adhere to
3. Each observation is independent
 - We'll violate this one for the sake of *causality*
4. Homoskedasticity: Variance in errors is constant
 - This is important
5. Not too much multicollinearity
 - Each \hat{x}_i should be relatively independent from the others
 - Some is OK

Linear model implementation

What exactly is a linear model?

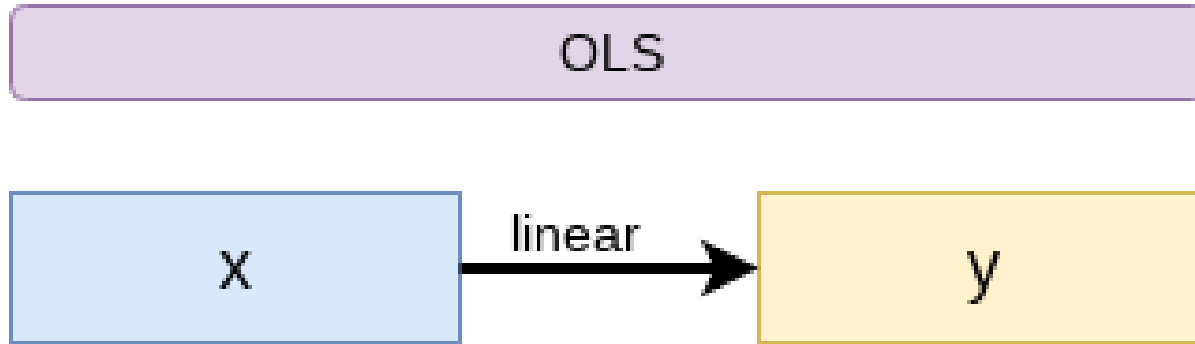
- Anything OLS is linear
- Many transformations can be recast to linear
 - Ex.: $\log(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 \cdot x_2$
 - This is the same as $y' = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ where:
 - $y' = \log(y)$
 - $x_3 = x_1^2$
 - $x_4 = x_1 \cdot x_2$

Linear models are *very* flexible



source: wikipedia

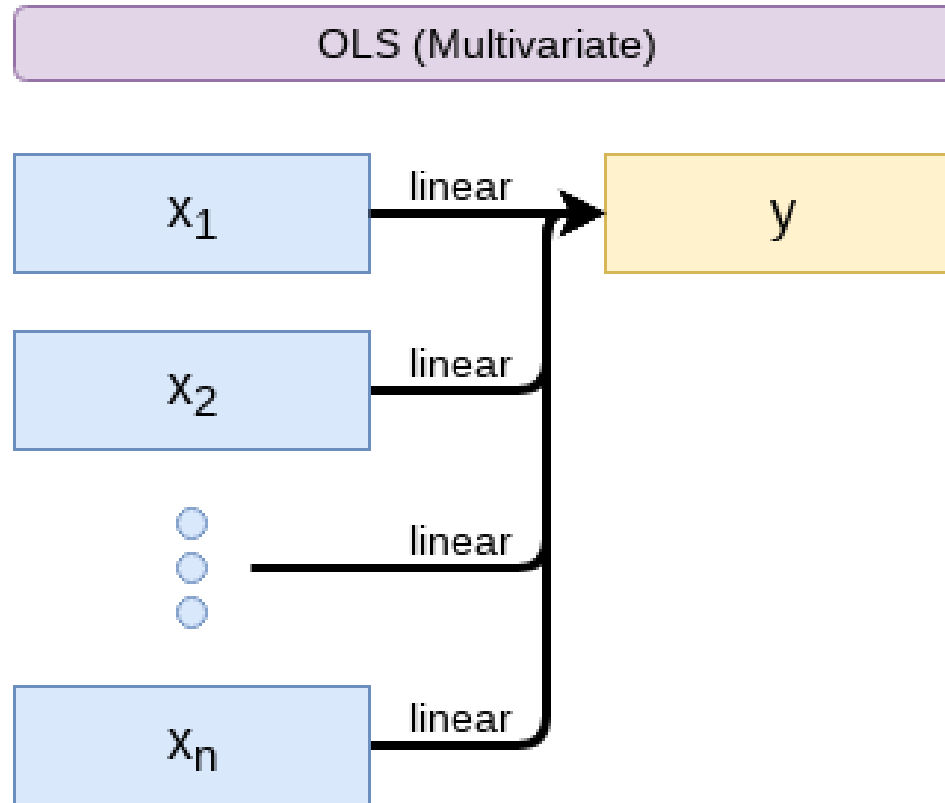
Mental model of OLS: 1 input



Simple OLS measures a simple linear relationship between an input and an output

- e.g.: Future revenue regressed on assets

Multiple inputs



OLS measures simple linear relationships between a set of inputs and one output

- e.g.: Future revenue regressed on multiple accounting and macro variables

Model selection

■ We will introduce many models. Pick what fits your problem!

- For forecasting a quantity
 - Usually some sort of linear model regressed using OLS
- For forecasting a binary outcome
 - Usually logit or a related model
- For forensics:
 - Usually logit or a related model

■ automated model selection



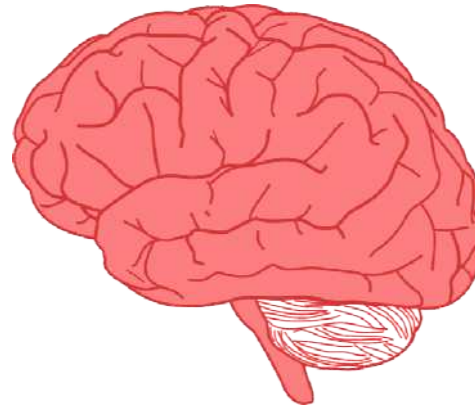
Variable selection

| Feature engineering

- The options:
 1. Use your own knowledge to select variables
 2. Use a selection model to automate it

Own knowledge

- Build a model based on your knowledge of the problem and situation
- This is generally better
 - The result should be more interpretable
 - For prediction, you should know relationships better than most algorithms



Automated variable selection

- Traditional methods include:
 - Forward selection: Start with nothing and add variables with the most contribution to $\text{Adj } R^2$ until it stops going up
 - Backward selection: Start with all inputs and remove variables with the worst (negative) contribution to $\text{Adj } R^2$ until it stops going up
 - Stepwise selection: Like forward selection, but drops non-significant predictors
- Newer methods:
 - Lasso and Elastic Net based models
 - Optimize with high penalties for complexity (i.e., # of inputs)
- We will discuss these in future sessions

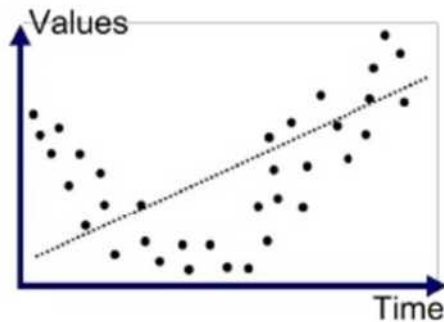


The overfitting problem

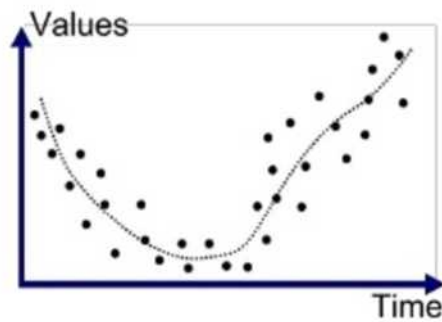
Or: Why do we like simpler models so much?

- Overfitting happens when a model fits in-sample data *too well*...
 - To the point where it also models any idiosyncrasies or errors in the data
 - This harms prediction performance
 - Directly harming our forecasts

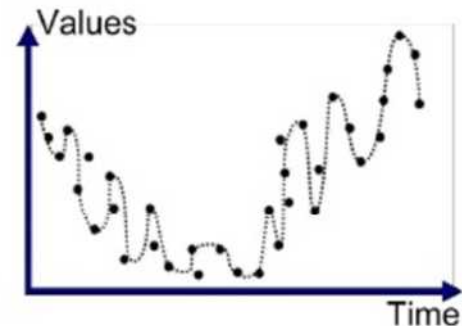
An overfitted model works really well on its own data, and quite poorly on new data



Underfitted



Good Fit/Robust

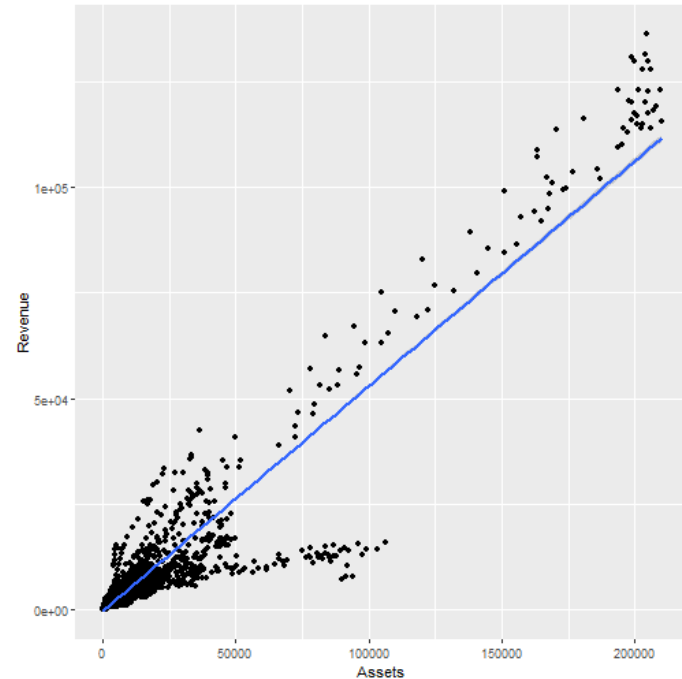


Overfitted

Statistical tests and Interpretation

Coefficients

- In OLS: β_i
- A change in x_i by 1 unit leads to a change in y by β_i
- Essentially, the slope between x and y
- The blue line in the chart is the regression line for $\hat{Revenue} = \alpha + \beta_i \hat{Assets}$ for retail firms since 1960



P-values

- p -values tell us the probability that an individual result is due to random chance

"The P value is defined as the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed."

-- Dahiru 2008

- These are very useful, particularly for a frequentist approach
- First used in the 1700s, but popularized by Ronald Fisher in the 1920s and 1930s
- If $p < 0.05$ and the coefficient matches our mental model, we can consider this as supporting our model (i.e. rejecting the null)
 - If $p < 0.05$ but the coefficient is opposite, then it is suggesting a problem with our model
 - If $p > 0.10$, it is rejecting the alternative hypothesis
- If $0.05 < p < 0.10$ it depends...
 - For a small dataset or a complex problem, we can use 0.10 as a cutoff
 - For a huge dataset or a simple problem, we should use 0.05

R-square

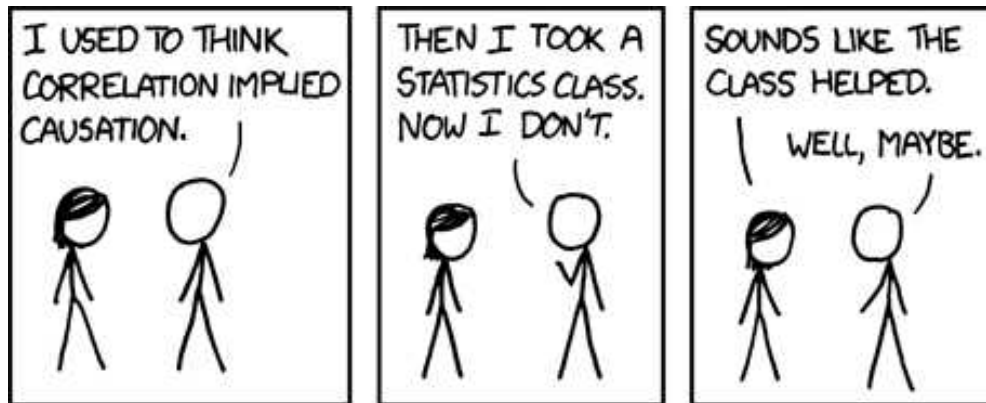
- $R^2 = \text{Explained variation} / \text{Total variation}$
 - Variation = difference in the observed output variable from its own mean
- A high R^2 indicates that the model fits the data very well
- A low R^2 indicates that the model is missing much of the variation in the output
- R^2 is technically a *biased* estimator
 - more independent variables, higher R^2
- Adjusted R^2 downweights R^2 and makes it unbiased
 - $R^2_{Adj} = P * R^2 + 1 - P$
 - Where $P = \frac{n-1}{n-p-1}$
 - n is the number of observations
 - p is the number of inputs in the model

Causality

What is causality?

$A \rightarrow B$

- Causality is *A causing B*
 - This means more than *A* and *B* are correlated
- i.e., If *A* changes, *B* changes. But *B* changing doesn't mean *A* changed
 - Unless *B* is 100% driven by *A*
- Very difficult to determine, particularly for events that happen [almost] simultaneously
- Examples of correlations that aren't causation



Time and causality

$A \rightarrow B$ or $A \leftarrow B$?

$A_t \rightarrow B_{t+1}$

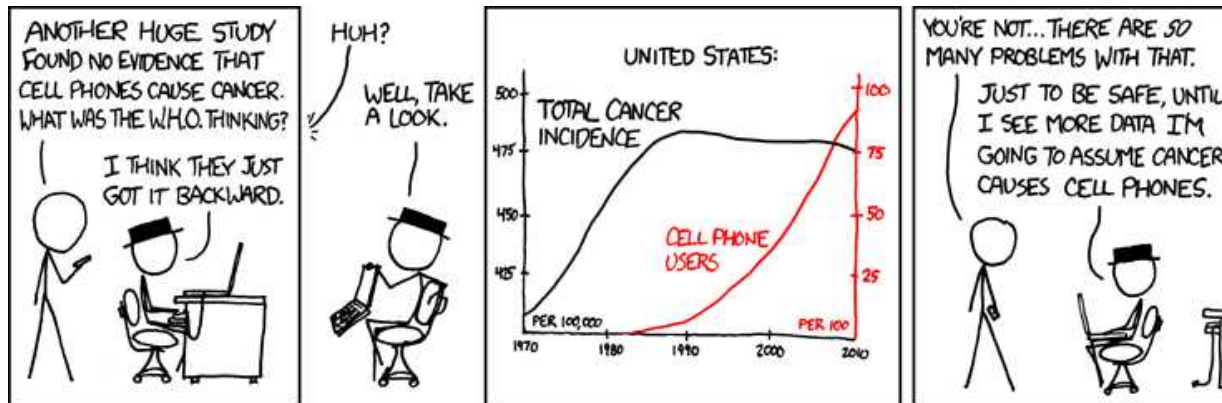
- If there is a separation in time, it's easier to say A caused B
 - Observe A , then see if B changes after
- Conveniently, we have this structure when forecasting
 - e.g.:

$$Revenue_{t+1} = Revenue_t + \dots$$

Time and causality break down

$A_t \rightarrow B_{t+1}$? OR $C \rightarrow A_t$ and $C \rightarrow B_{t+1}$?

- The above illustrates the *Correlated omitted variable problem*
 - A doesn't cause B ... Instead, some other force C causes both
 - Bane of social scientists everywhere
- This is less important for predictive analytics, as we care more about performance, but...
 - It can complicate interpreting your results
 - Figuring out C can help improve your model's predictions
 - So find C !

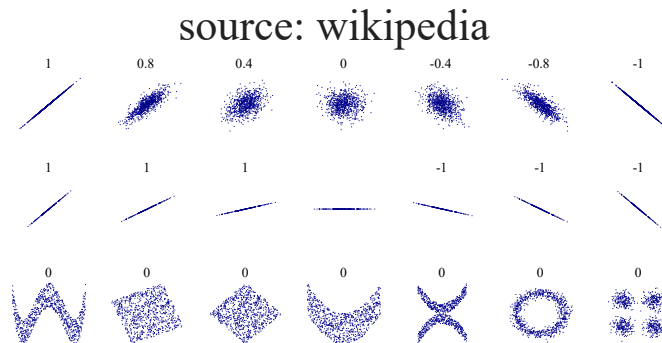


Discussion

Some executives believe that all they need to do is establish correlation. Wrong!

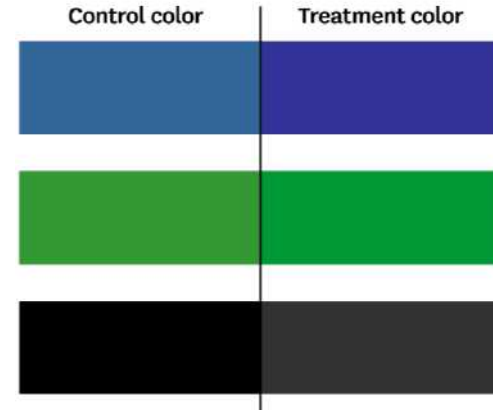
- **A/B Test: The Surprising Power of Online Experiments**
- Further reading: **Causal Inference cheat sheet for data scientists**

so does causation imply correlation?



Small Changes with a Huge Impact

Bing's experiments showed that slightly darker blues and greens in titles and a slightly lighter black in captions improved the users' experience. When rolled out to all users, the color changes boosted revenue by more than \$10 million annually.



FROM "THE SURPRISING POWER OF ONLINE EXPERIMENTS," SEPTEMBER-OCTOBER 2017, BY RON KOHAVI AND STEFAN THOMKE

© HBR.ORG

Summary of Session 1

For next week

- start your "data analyst with R" career track on Datacamp
- Review statistics foundation
- Pick a book on R and study it, such as **R Cookbook** or **R for Data Science**
- Install **R** and **RStudio** if you have not done so

R packages used in this slide

This slide was prepared on 2021-08-22 from Session_1s.Rmd with R version 4.0.3
(2020-10-10) Bunny-Wunnies Freak Out on Windows 10 x64 build 18362 ☺.

The attached packages used in this slide are:

##	forcats	stringr	dplyr	purrr	readr	tidyr	tibble
##	"0.5.1"	"1.4.0"	"1.0.4"	"0.3.4"	"1.4.0"	"1.1.2"	"3.0.6"
##	tidyverse	ggplot2	ngramr	kableExtra	knitr		
##	"1.3.0"	"3.3.3"	"1.7.2"	"1.1.0"	"1.33"		