Appeared in: Statistica Sinica, 2007, 17, 817-837.

# STATISTICS WITH ESTIMATED PARAMETERS[1]

BY Z. L. YANG, Y. K. TSE

*School of Economics and Social Sciences, Singapore Management University,*

*90 Stamford Road, Singapore 178903.*

zlyang@smu.edu.sg, yktse@smu.ed.sg

Z. D. BAI

*KLASMOE & School of Mathematics and Statistics, Northeast Normal University,*

*Changchun 130024, Jilin, China*

*and*

*Dept. of Statistics and Applied Probability, National University of Singapore,*

*6 Science Drive 2, Singapore 117543.*

stabaizd@nus.edu.sg

September 2005

*Abstract:* This paper studies a general problem of making inferences for functions of two sets of parameters where, when the first set is given, there exists a statistic with a known distribution. We study the distribution of this statistic when the first set of parameters is unknown and is replaced by an estimator. We show that under mild conditions the variance of the statistic is inflated when the unconstrained maximum likelihood estimator (MLE) is used, but deflated when the constrained MLE is used. The results are shown to be useful in hypothesis testing and confidence-interval construction in providing simpler and improved inference methods than do the standard large sample likelihood inference theories. We provide three applications of our theories, namely Box-Cox regression, dynamic regression,

[1]We thank Peter Robinson, Anil Bera, W. K. Li and a referee for their helpful comments that have led to improvements in this paper. Thanks are also due to seminar participants at the University of Hong Kong, the Chinese University of Hong Kong, and the 6th ICSA International Conference for their useful comments. The research support from the Wharton-SMU research center, Singapore Management University, is gratefully acknowledged.

Appeared in: Statistica Sinica, 2007, 17, 817-837.

and spatial regression, to illustrate the generality and versatility of our results.

*Key words and phrases:* Asymptotic distribution, Finite sample performance, Index parameter, Variance deflation, Variance inflation.

Appeared in: Statistica Sinica, 2007, 17, 817-837.

## 1. Introduction

In a variety of econometric problems, the models for the data $\mathbf{y} = \{y_1, y_2, \cdots, y_n\}$ often involve two sets of parameters: $\theta$ and $\lambda$. A distinct feature is that when $\lambda$ is known, model inferences are simple. Examples include the following: i) Weibull duration analysis, where knowing the shape parameter reduces the model to an exponential; ii) Box-Cox regression, where knowing the transformation parameter reduces the model to standard linear regression; iii) dynamic regression, where knowing the coefficients of the lagged dependent variables and the serial correlation coefficients reduces the model to weighted linear regression; and iv) spatial regression, where knowing the coefficients of spatial effects reduces the model to either standard or weighted linear regression. We call $\lambda$ the vector of *index parameters*. In each of the above examples, exact inference methods are usually available when $\lambda$ is given.

When $\lambda$ is unknown, a naive approach is to conduct model inference by substituting an estimate $\tilde{\lambda}$ for $\lambda$ in the inferential statistic. To clarify the idea, we consider first a simple case where inference concerns the parameter $\theta$. Suppose there is a statistic $T(\mathbf{y}; \lambda, \theta)$ with a known distribution, so that inference for $\theta$ can be conducted when $\lambda$ is known. When $\lambda$ is unknown, it is replaced by $\tilde{\lambda}$ to give $T(\mathbf{y}; \tilde{\lambda}, \theta)$. This raises the following questions. What is the distribution of $T(\mathbf{y}; \tilde{\lambda}, \theta)$? How does one adjust $T(\mathbf{y}; \tilde{\lambda}, \theta)$ so as to allow inference for $\theta$ to proceed in the same manner as when $\lambda$ is known? Some related questions include the following. Which estimator of $\lambda$ should one use: constrained (for given $\theta$) or unconstrained? Does it make a difference?

This paper develops general theories to deal with these issues in the broader set up where inference concerns a general (vector-valued) parametric function $\psi = g(\lambda, \theta)$. We show that, under mild conditions, the asymptotic variance of $T(\mathbf{y}; \tilde{\lambda}, \psi)$ is inflated over that of $T(\mathbf{y}; \lambda, \psi)$ when $\tilde{\lambda}$ is the unconstrained estimator of $\lambda$, but deflated when $\tilde{\lambda}$ is the constrained estimator for a given $\theta$. In either case, the standardized statistic using the correct asymptotic variance can be used for inference. More importantly, when the finite sample distribution of $T(\mathbf{y}; \lambda, \psi)$ is known, $T(\mathbf{y}; \tilde{\lambda}, \psi)$ can be corrected

Appeared in: Statistica Sinica, 2007, 17, 817-837.

using the exact first and second moments of $T(\mathbf{y};\lambda,\psi)$, along with the variance inflation/deflation factor. Then, referring the corrected $T(\mathbf{y};\tilde{\lambda},\psi)$ to the exact distribution of $T(\mathbf{y};\lambda,\psi)$ gives procedures with an improved finite sample performance.

Our approach is related to the delta method and the likelihood ratio method, but with clear distinctions: our approach is simpler and is able to take advantage of exact inference methods when $\lambda$ is given, resulting in inferences with an improved finite sample behavior.

The rest of the paper is organized as follows. Section 2 presents an example to further motivate our ideas and to shed light on the type of results we are expecting. Section 3 presents the main results. Section 4 contains three applications of the theorems.

## 2. A MOTIVATING EXAMPLE: THE WEIBULL DURATION MODEL

For illustrative purpose, we consider the simple situation where $y_1, y_2, \cdots, y_n$ are independent and identically distributed (iid) Weibull random variables with probability density function $\lambda\theta^{-\lambda}y^{\lambda-1}\exp[-(y/\theta)^{\lambda}]$, $\lambda > 0$. The Weibull distribution is one of the most popular models for modeling economic durations (Kiefer, 1988).

**The simple set-up.** We first consider inference for the scale parameter $\theta$, when $\lambda$ is treated as the index parameter. Define

$$T(\mathbf{y};\lambda,\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\theta}\right)^{\lambda} - 1.$$

Then $\sqrt{n}\,T(\mathbf{y};\lambda,\theta) \xrightarrow{D} N(0,1)$ and the finite sample distribution of $2\sum_{i=1}^{n}(y_i/\theta)^{\lambda}$ is chi-squared with $2n$ degrees of freedom. Thus, if $\lambda$ is known, exact inference about $\theta$ can be conducted based on $2\sum_{i=1}^{n}(y_i/\theta)^{\lambda}$.

Denote the unconstrained maximum likelihood estimator (MLE) by $\hat{\lambda}$ and the constrained (for a given $\theta$) MLE of $\lambda$ by $\hat{\lambda}_\theta$. Define

$$T(\mathbf{y};\hat{\lambda},\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\theta}\right)^{\hat{\lambda}} - 1,$$

$$T(\mathbf{y};\hat{\lambda}_\theta,\theta)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\theta}\right)^{\hat{\lambda}_\theta}-1.$$

Here, the standard asymptotic results of the maximum likelihood theory apply. Furthermore, as shown in Appendix A.1, $\sqrt{n}\,T(\mathbf{y};\lambda,\theta)$ and $\sqrt{n}\,(\hat{\lambda}-\lambda)$ are asymptotically independent with

$$\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\theta)\xrightarrow{D}N(0,1+c_1^2), \tag{1}$$

and $\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\theta,\theta)$ is asymptotically independent of $\sqrt{n}\,(\hat{\lambda}_\theta-\lambda)$ with

$$\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\theta,\theta)\xrightarrow{D}N(0,1-c_2^2), \tag{2}$$

where $c_1^2=6(1-\gamma)^2/\pi^2=0.1087$, $c_2^2=(1-\gamma)^2/[(1-\gamma)^2+(\pi^2/6)]=0.0980$, and $\gamma=0.5772$ is Euler's constant. As we shall see, these results can also be obtained as direct applications of Theorems 1 and 2 below. Hence, the use of $\hat{\lambda}$ inflates the asymptotic variance of $T$, whereas the use of $\hat{\lambda}_\theta$ deflates the asymptotic variance. In either case, it is very easy to adjust the statistic to give a limiting standard normal distribution. In particular, the statistics

$$T^*(\mathbf{y};\hat{\lambda},\theta)=\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\theta)/\sqrt{1+c_1^2}\quad\text{and}\quad T^*(\mathbf{y};\hat{\lambda}_\theta,\theta)=\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\theta,\theta)/\sqrt{1-c_2^2}$$

are both asymptotically $N(0,1)$. To test $H_0:\theta=\theta_0$, both adjusted statistics are equally simple to use, but to construct a confidence interval for $\theta$, it is simpler to use the former. Specifically, a two-sided $100(1-\alpha)\%$ large sample confidence interval (CI) for $\theta$ based on $T^*(\mathbf{y};\hat{\lambda},\theta)$ takes the following explicit form

$$\left\{\left(\frac{n^{-1/2}\sum_{i=1}^n y_i^{\hat{\lambda}}}{\sqrt{n}+z_{\alpha/2}\sqrt{1+c_1^2}}\right)^{\frac{1}{\hat{\lambda}}},\ \left(\frac{n^{-1/2}\sum_{i=1}^n y_i^{\hat{\lambda}}}{\sqrt{n}-z_{\alpha/2}\sqrt{1+c_1^2}}\right)^{\frac{1}{\hat{\lambda}}}\right\}, \tag{3}$$

whereas the same interval based on $T^*(\mathbf{y};\hat{\lambda}_\theta,\theta)$ is defined implicitly by the set

$$\left\{\theta:-z_{\alpha/2}\le T^*(\mathbf{y};\hat{\lambda}_\theta,\theta)\le z_{\alpha/2}\right\}, \tag{4}$$

which has to be solved numerically.

**The general set-up.** Suppose now inference concerns $\psi \equiv g(\lambda, \theta)$, a smooth function of both parameters, representing a survivor-related quantity such as i) the $k$th moment, where $g(\lambda, \theta) = \theta^k \Gamma(1 + k/\lambda)$, ii) the survivor function, where $g(\lambda, \theta) = \exp[-(y/\theta)^\lambda]$, iii) the hazard function, where $g(\lambda, \theta) = \theta^{-\lambda} \lambda y^{\lambda-1}$, and iv) the $p$-quantile, where $g(\lambda, \theta) = \theta[1 - \log(1-p)]^{1/\lambda}$. Note that $y$ in ii) and iii) is a given constant. Suppose the function $g$ is invertible in $\theta$ as are the cases above. Let $\theta = g^{-1}(\lambda, \psi) \equiv f(\lambda, \psi)$. Define

$$T(\mathbf{y}; \lambda, \psi) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i}{f(\lambda, \psi)} \right)^{\lambda} - 1,$$

which is $T(\mathbf{y}; \lambda, \theta)$ given above, and hence has a limiting standard normal distribution. The difference is that $T(\mathbf{y}; \lambda, \psi)$ is now considered as the statistic used for inference concerning $\psi$ and that $\lambda$, the index parameter to be substituted, also appears in the function $f$. When $\lambda$ is unknown, define

$$T(\mathbf{y}; \hat{\lambda}, \psi) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i}{f(\hat{\lambda}, \psi)} \right)^{\hat{\lambda}} - 1,$$

$$T(\mathbf{y}; \hat{\lambda}_\psi, \psi) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i}{f(\hat{\lambda}_\psi, \psi)} \right)^{\hat{\lambda}_\psi} - 1,$$

where $\hat{\lambda}$ is the unconstrained MLE of $\lambda$, and $\hat{\lambda}_\psi$ is the constrained MLE of $\lambda$ for a given $\psi$.

It can be shown (see Appendix A.1) that $\sqrt{n}\, T(\mathbf{y}; \lambda, \psi)$ and $\sqrt{n}\, (\hat{\lambda} - \lambda)$ are asymptotically independent with

$$\sqrt{n}\, T(\mathbf{y}; \hat{\lambda}, \psi) \xrightarrow{D} N(0, 1 + c_1^2(\lambda, \psi)), \tag{5}$$

and that $\sqrt{n}\, T(\mathbf{y}; \hat{\lambda}_\psi, \psi)$ and $\sqrt{n}\, (\hat{\lambda}_\psi - \lambda)$ are asymptotically independent with

$$\sqrt{n}\, T(\mathbf{y}; \hat{\lambda}_\psi, \psi) \xrightarrow{D} N(0, 1 - c_2^2(\lambda, \psi)), \tag{6}$$

where

$$\begin{aligned} c_1^2(\lambda, \psi) = &\ \tfrac{6}{\pi} \left( 1 - \gamma + \tfrac{\lambda^2 f_\lambda(\lambda, \psi)}{f(\lambda, \psi)} \right), \\ c_2^2(\lambda, \psi) = &\ \left( \tfrac{1-\gamma}{\lambda} - \tfrac{\lambda f_\lambda(\lambda, \psi)}{f(\lambda, \psi)} \right)^2 \left( \tfrac{\pi^2}{6\lambda^2} + \left( \tfrac{1-\gamma}{\lambda} - \tfrac{\lambda f_\lambda(\lambda, \psi)}{f(\lambda, \psi)} \right)^2 \right)^{-1}. \end{aligned}$$

Appeared in: Statistica Sinica, 2007, 17, 817-837.

Thus, as before, the use of the unconstrained estimator results in variance inflation, and the use of the constrained estimator results in variance deflation. Both statistics can easily be adjusted to give asymptotic $N(0,1)$ statistics

$$T^*(\mathbf{y};\hat{\lambda},\psi) = \frac{\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\psi)}{\sqrt{1+c_1^2(\hat{\lambda},\hat{\psi})}} \quad \text{and} \quad T^*(\mathbf{y};\hat{\lambda}_\psi,\psi) = \frac{\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\psi,\psi)}{\sqrt{1-c_2^2(\hat{\lambda}_\psi,\psi)}},$$

which can be used to conduct inferences for $\psi$. In testing $H_0:\psi=\psi_0$, both statistics are equally simple to use, but to construct confidence interval, it is again simpler to use the statistic based on $\hat{\lambda}$, which gives a two-sided $100(1-\alpha)\%$ large-sample CI for $\psi$ with explicit lower and upper limits as follows:

$$\{g[\hat{\lambda}, L(\hat{\lambda})], \{g[\hat{\lambda}, U(\hat{\lambda})]\} \tag{7}$$

where $L(\hat{\lambda})$ and $U(\hat{\lambda})$ are the lower and upper confidence limits given in (3), with $c_1$ replaced by $c_1(\hat{\lambda},\hat{\psi})$. Using the statistic with $\hat{\lambda}_\psi$ substituting for $\lambda$, however, the CI for $\psi$ has to be defined implicitly in a similar way as in (4), i.e.,

$$\left\{\theta : -z_{\alpha/2} \le T^*(\mathbf{y};\hat{\lambda}_\psi,\psi) \le z_{\alpha/2}\right\}, \tag{8}$$

which again has to be solved through numerical iterations.

**The improved inferences.** The idea of *simplicity* is clearly illustrated by the above developments: implementation of suggested methods does not need the calculation of the information matrix whereas the implementation of the delta method does. Furthermore, in CI constructions, it is simpler to use $T^*(\mathbf{y};\hat{\lambda},\psi)$ than $T^*(\mathbf{y};\hat{\lambda}_\psi,\psi)$ as the former does not involve numerical iterations. We now illustrate the idea that the above developments also lead to improved finite sample inferences.

Note that $\sqrt{n}T(\mathbf{y};\lambda,\theta) = \frac{1}{2\sqrt{n}}\chi^2_{2n} - \sqrt{n}$, where $\chi^2_{2n}$ is a chi-squared random variable with $2n$ degrees of freedom. As the exact mean and variance of $\sqrt{n}T(\mathbf{y};\lambda,\theta)$ are 0 and 1, all the $T^*$ statistics defined above do not need to be further adjusted (otherwise they do as explained in Sec. 3.4). However, finite sample performance of inference procedures based on these $T^*$ statistics will be improved if they are referred to the distributions of $\frac{1}{2\sqrt{n}}\chi^2_{2n} - \sqrt{n}$ instead of $N(0,1)$. In particular, in the CIs defined in (3), (4), (7) and (8), replace $-z_{\alpha/2}$ by $\frac{1}{2\sqrt{n}}\chi^2_{2n}(1-\frac{\alpha}{2}) - \sqrt{n}$, and $z_{\alpha/2}$ by $\frac{1}{2\sqrt{n}}\chi^2_{2n}(\frac{\alpha}{2}) - \sqrt{n}$, where $\chi^2_{2n}(1-\frac{\alpha}{2})$ and $\chi^2_{2n}(\frac{\alpha}{2})$ are, respectively, the lower and upper $100(\alpha/2)\%$ points of the $\chi^2_{2n}$ distribution. Monte Carlo results confirming the better finite sample performance of this CI are available from the first author upon request.

The above example clearly illustrates the simplicity and improved finite sample properties of the proposed inference methods. It shows that i) the effect of estimating the index parameter cannot be ignored; ii) the variance of the inferential statistic is inflated when using the unconstrained estimator, but deflated when using the constrained estimator; iii) in either case, the statistic can be easily adjusted to account for index parameter estimation; and iv) both adjusted statistics provide simple tests for the parameter of interest, but only the statistic with unconstrained estimator provides explicit solutions for confidence interval construction. To our knowledge, the results above (equations (1), (2), (5) and (6)) are new. They can be applied to economic duration analysis and are extendable to the case of censored data. Motivated by this example, we provide some general results in the next section.

## 3. The main results

We give a general treatment of the problem by considering the parameter of interest to be $\psi = g(\lambda, \theta)$, a general smooth function of all parameters. Interesting special cases include i) all the elements of $\theta$ are the parameters of interest, i.e., $\psi = \theta$, ii) some elements of $\theta$ are the parameters of interest, i.e., $\psi = \theta_1$, and iii) there are no parameters of interest, i.e., $\psi$ is an empty vector, as in goodness of fit tests and residual-based diagnostics. It is desired to find the limiting distribution of $T(\mathbf{y}; \tilde{\lambda}, \psi)$ with $\tilde{\lambda}$ denoting a general estimator of $\lambda$ that may be the constrained MLE given $\psi$, or the unconstrained MLE. As in the earlier example about the different behaviors of $T(\mathbf{y}; \tilde{\lambda}, \psi)$ when using unconstrained or constrained estimator for $\lambda$, we treat the two cases separately.

Appeared in: Statistica Sinica, 2007, 17, 817-837.

### 3.1. Assumptions and Preliminaries

Throughout the paper, we assume that the usual regularity conditions for maximum likelihood (ML) estimation holds (See, for example, Godfrey (1988); Davidson and MacKinnon (1993)). We also assume that $\sqrt{n}\,T(\mathbf{y};\lambda,\psi)$ follows exactly, or asymptotically, a normal distribution with mean vector zero and variance-covariance matrix $V$, where $V$ may be parameter dependent but can be consistently estimated. Also, $V$ is nonsingular (for variance correction purposes). As our formulation starts with the case of a known $\lambda$, $T$ should have the same dimension as $\psi$ (one needs as many equations as unknowns) and $\psi$ should have a dimension less than or equal to $\theta$.

We denote the log likelihood function by $L(\lambda,\theta)$, the score function by $U(\lambda,\theta)$ and the Fisher information matrix by $I(\lambda,\theta)$. Write $U(\lambda,\theta)=(U_\lambda(\lambda,\theta)',U_\theta(\lambda,\theta)')'=(\partial L(\lambda,\theta)/\partial\lambda',\partial L(\lambda,\theta)/\partial\theta')'$. Let

$$A=\lim_{n\to\infty}\left(\frac{1}{n}I(\lambda,\theta)\right),$$

which is partitioned according to $(\lambda,\theta)$ into sub-blocks $A_{ij}$, $i,j=1,2$. The following is a generic assumption that applies to both the unconstrained and constrained cases.

**Assumption I:** *There is a matrix* $B=\lim_{n\to\infty}\mathrm{E}[\partial T(\mathbf{y};\lambda,\theta)/\partial\lambda']$, *such that*

$$\sqrt{n}\,T(\mathbf{y};\tilde\lambda,\psi)=\sqrt{n}\,T(\mathbf{y};\lambda,\psi)+B\sqrt{n}\,(\tilde\lambda-\lambda)+o_p(1),$$

*for any consistent estimator* $\tilde\lambda$ *of* $\lambda$.

This assumption ensures the validity of the Taylor expansion of $\sqrt{n}\,T(\mathbf{y};\tilde\lambda,\psi)$. It is clearly not restrictive. In certain situations such as when $\tilde\lambda$ is the unconstrained estimator, the condition on $B$ can be further relaxed to require $T$ to be only asymptotically smooth.

### 3.2. Substituting the Unconstrained Estimator

Appeared in: Statistica Sinica, 2007, 17, 817-837.

**Assumption II:** *$T(\mathbf{y};\lambda,\psi) = k(n^{-\frac{1}{2}}U_\theta(\lambda,\theta)) + o_p(1)$, where $k$ is a measurable function of $U_\theta(\lambda,\theta)$, the score component corresponding to $\theta$.*

Assumption II holds for the score statistic, and hence for the Wald as well as the likelihood ratio statistics as the latter two are asymptotically equivalent to the score statistic (see Godfrey (1988); Cox and Hinkley (1974)). An intuitive interpretation of this assumption is as follows. When $\lambda$ is known, to make inference about $\psi$, one has to estimate the model (the parameters $\theta$) by solving the first-order conditions $U_\theta(\lambda,\theta) = 0$. Then, one sets up the statistic based on this estimating equation for testing and confidence interval construction for $\psi$. As a result, the statistic becomes a measurable function of $U_\theta(\lambda,\theta)$ or an asymptotically equivalent version of it.

**Lemma 1:** *Under the usual regularity conditions of ML estimation, $n^{-\frac{1}{2}}U_\theta(\lambda,\theta)$ and $\sqrt{n}(\hat{\lambda}-\lambda)$ are asymptotically independent.*

The result of Lemma 1 depends critically on the information equality. It says that in the ML estimation framework the conditional estimation of $\theta$ (given $\lambda$) is asymptotically independent of the unconditional estimation of $\lambda$. As shown below, this lemma leads to an important result regarding the limiting distribution of $T(\mathbf{y};\hat{\lambda},\psi)$. From Lemma 1, it follows immediately that $\hat{\theta}_\lambda$ (the constrained estimator of $\theta$ given $\lambda$) is asymptotically independent of $\hat{\lambda}$.

**Theorem 1:** *Under the usual regularity conditions of ML estimation and Assumptions I and II, $\sqrt{n}T(\mathbf{y};\lambda,\psi)$ and $\hat{\lambda}$ are asymptotically independent, and*

$$\sqrt{n}T(\mathbf{y};\hat{\lambda},\psi) \xrightarrow{D} N(0, V + BA_{11.2}^{-1}B'),$$

*where $A_{11.2}^{-1} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$ is the upper-left-corner block of $A^{-1}$.*

Theorem 1 says that the consequence of using the unconstrained estimator $\hat{\lambda}$ in place of $\lambda$ causes the variance of $T$ to be inflated. With a simple adjustment to the variance of $T$, Theorem 1 allows inference about $\psi$ to be carried out in the same manner as when $\lambda$ is known, provided the *variance inflation factor*, $BA_{11.2}^{-1}B'$, can be consistently estimated. This is often an easy task since $B$ can be consistently estimated by $(\partial/\partial\lambda)T(\mathbf{y};\hat{\lambda},\hat{\psi})$, and $A_{11.2}^{-1}$, the asymptotic variance of $\sqrt{n}(\hat{\lambda}-\lambda)$, can be consistently estimated by $nI_{11.2}^{-1}(\hat{\lambda},\hat{\psi})$, or by $n\tilde{I}_{11.2}^{-1}(\hat{\lambda},\hat{\psi})$, or simply by

$$-n\left(\frac{\partial^2}{\partial\lambda\partial\lambda'}L_{\max}(\hat{\lambda})\right)^{-1}, \tag{9}$$

where $\tilde{I}$ is the observed information matrix, $L_{\max}(\lambda)$ is the concentrated or partially maximized (over $\theta$) log likelihood of $\lambda$, and $I_{11.2}$ and $\tilde{I}_{11.2}$ are defined in the same way as $A_{11.2}$. The last method is particularly simple as the second derivative can be calculated numerically (see Carroll and Ruppert (1988, p129)). It makes the application of Theorem 1 more handy when a simple expression for the concentrated log likelihood for $\lambda$ is available (see the applications in Section 4).

### 3.3. Substituting the Constrained Estimator

The case of substituting the constrained MLE $\hat{\lambda}_\psi$ (given $\psi$) for $\lambda$ is more complex. The reason is that $\hat{\lambda}_\psi$ is subject to constraints imposed on the parameters through $H_0 : \psi = \psi_0$. To overcome this difficulty, we reparameterize the model by defining a one-to-one transformation: $(\lambda,\theta) \longleftrightarrow (\lambda,\psi,\phi)$, where $\psi \equiv g(\lambda,\theta)$ is of dimension less than or equal to that of $\theta$, $\phi$ represents (loosely speaking) the remaining components of $\theta$, and $\lambda$ remains the index parameter. In this context, $\phi$ is the vector of nuisance parameters. Note that $\lambda$ has to be estimated jointly with $\phi$. Thus, the estimation process involves only the score functions for $\lambda$ and $\phi$. As $\psi$ is specified under the null hypothesis, it is suppressed from the notation. Denote the scores for $\lambda$ and $\phi$ by $U_\lambda^\circ(\lambda,\phi)$ and $U_\phi^\circ(\lambda,\phi)$ and the corresponding information sub-matrix by $I^\circ(\lambda,\phi)$, where the superscript $^\circ$ indicates that the corresponding quantity is obtained under reparameterization and constrained estimation. Let $A^\circ = \lim_{n\to\infty}\left(\frac{1}{n}I^\circ(\lambda,\phi)\right)$.

**Assumption III:** *For every $\lambda$, there is joint convergence in law to normality:*

$$\begin{bmatrix} \sqrt{n}\,T(\mathbf{y};\lambda,\psi) \\ \sqrt{n}\,(\hat{\lambda}_\psi-\lambda) \end{bmatrix} \xrightarrow{D} N\left[0, \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}\right].$$

Appeared in: Statistica Sinica, 2007, 17, 817-837.

*where* $V_{11} = V$*, and* $V_{22} = A^{\circ -1}_{11.2}$*, the upper-left-corner block of* $A^{\circ -1}$*. The dispersion matrix is assumed to be nonsingular.*

**Assumption IV:** $\mathrm{E}[T(\mathbf{y}; \lambda, \psi)]$ *is free of* $\lambda$*, and there is an integrable function* $h(\mathbf{y})$ *such that, in a neighborhood of* $(\lambda, \psi, \phi)$*, the following conditions hold:*

$$
\begin{array}{ll}
(a) & |[\frac{\partial}{\partial\lambda}T(\mathbf{y};\lambda,\psi)]p(\mathbf{y};\lambda,\psi,\phi)| \le h(\mathbf{y}), \quad |T(\mathbf{y};\lambda,\psi)[\frac{\partial}{\partial\lambda}p(\mathbf{y};\lambda,\psi,\phi)]| \le h(\mathbf{y}), \\
(b) & |[\frac{\partial}{\partial\phi}T(\mathbf{y};\lambda,\psi)]p(\mathbf{y};\lambda,\psi,\phi)| \le h(\mathbf{y}), \quad |T(\mathbf{y};\lambda,\psi)[\frac{\partial}{\partial\phi}p(\mathbf{y};\lambda,\psi,\phi)]| \le h(\mathbf{y}),
\end{array}
$$

*where* $p(\mathbf{y}; \lambda, \psi, \phi)$ *is the joint probability density function of* $\mathbf{y}$*.*

Note that the first part of (b) in Assumption IV always holds as the quantity is actually zero due to the fact that $T(\mathbf{y}; \lambda, \psi)$ is free of $\phi$.

**Theorem 2:** *Under the usual regularity conditions of ML estimation and Assumptions I, III and IV,* $\sqrt{n}T(\mathbf{y}; \hat{\lambda}_\psi, \psi)$ *and* $\sqrt{n}(\hat{\lambda}_\psi - \lambda)$ *are asymptotically independent and* $\sqrt{n}T(\mathbf{y}; \hat{\lambda}_\psi, \psi) \xrightarrow{D} N(0, V - BA^{\circ -1}_{11.2}B')$*.*

Theorem 2 shows that using the constrained estimator $\hat{\lambda}_\psi$ in place of $\lambda$ causes the variance of the statistic to be deflated. The deflation factor $BA^{\circ -1}_{11.2}B'$ and the original variance-covariance matrix $V$ can be consistently estimated, resulting in a properly standardized statistic to be used for testing and confidence-interval construction for $\psi$. One of the key quantities in the estimation of the deflation factor is the information sub-matrix under the new parameterization, which can be found through the relationship $I^\circ(\lambda, \psi, \phi) = J(\lambda, \psi, \phi)I(\lambda, \theta)J'(\lambda, \psi, \phi)$, where $J(\lambda, \psi, \phi) = \partial(\lambda', \theta')/\partial(\lambda', \psi', \phi')'$. Then, the desired quantity $I^\circ(\lambda, \phi)$ is just the sub-matrix of $I^\circ(\lambda, \psi, \phi)$ without its second row and second column.

Certain special cases of Theorem 2 are worthy of mention. When $\psi = \theta$, $\phi$ disappears and the result of Theorem 2 reduces to $T(\mathbf{y}; \hat{\lambda}_\theta, \theta) \xrightarrow{D} N(0, V - BA^{-1}_{11}B')$, which is given in Pierce (1982). Bera and Zuo (1996) used this result to derive a test for ARCH models. Bera and Kim (2002) used it to obtain a test for constant correlation in a bivariate conditional heteroscedasticity model. See also Tse′s (2002)

Appeared in: Statistica Sinica, 2007, 17, 817-837.

application to residual-based diagnostics for univariate and multivariate conditional heteroscedasticity models.

When $\psi$ and $\theta$ are of the same dimension, $\phi$ disappears and the transformation: $(\lambda, \theta) \longleftrightarrow (\lambda, \psi)$ becomes one-to-one. The result of Theorem 2 thus becomes $T(\mathbf{y}; \hat{\lambda}_\psi, \psi) \xrightarrow{D} N(0, V - BA^{\circ}{}_{11}^{-1}B')$. To calculate $A_{11}^{\circ}$, note that $J(\lambda, \psi)$ has rows $(\mathbf{I}, f_\lambda)$ and $(\mathbf{0}, f_\psi)$, where $f_\lambda = \frac{\partial}{\partial\lambda} f(\lambda, \psi)$, $f_\psi = \frac{\partial}{\partial\psi} f(\lambda, \psi)$, $f(\lambda, \psi) = \theta = g^{-1}(\lambda, \psi)$, $\mathbf{I}$ is an identity matrix, and $\mathbf{0}$ is a rectangular matrix of zeros. The dimensions of $\mathbf{I}$ and $\mathbf{0}$ are implicitly defined. Thus, $I_{\lambda\lambda}^{\circ} = I_{\lambda\lambda} + 2f_\lambda I_{\theta\lambda} + f_\lambda I_{\theta\theta} f_\lambda'$, which leads to $A_{11}^{\circ}$.

Finally, for the case where $\psi$ has dimension less than $\theta$, write $\theta = (\theta_1', \theta_2')'$ with $\theta_1$ and $\psi$ having the same dimension. Define $\psi = g(\lambda, \theta_1, \theta_2)$, and $\phi = \theta_2$, so that $\theta_1 = g^{-1}(\lambda, \psi, \phi) \equiv f(\lambda, \psi, \phi)$. Then

$$J(\lambda, \psi, \phi) = \begin{pmatrix} \mathbf{I} & f_\lambda & \mathbf{0} \\ \mathbf{0} & f_\psi & \mathbf{0} \\ \mathbf{0} & f_\phi & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I}$ and $\mathbf{0}$ in different positions are of different dimensions. The desired quantity $I^{\circ}(\lambda, \phi)$ is the submatrix of $I^{\circ}(\lambda, \psi, \phi) = J(\lambda, \psi, \phi) I(\lambda, \theta_1, \theta_2) J'(\lambda, \psi, \phi)$ obtained by deleting the second row and second column .

## 3.4. Improved Finite Sample Inference

As illustrated in Section 2, finite sample performance of the proposed inference procedures can be improved by referring to the exact distribution of the $\lambda$-known statistic $\sqrt{n}T(\mathbf{y}; \lambda, \psi)$ if it exists. We now generalize this idea. Let $\mu_n$ and $V_n$ be the finite sample mean and variance of $\sqrt{n}T(\mathbf{y}; \lambda, \psi)$. Clearly, $\mu_n \to 0$ and $V_n \to V$ as $n \to \infty$. Consider first the case of unconstrained substitution. Under Assumption I, we have $\sqrt{n}T(\mathbf{y}; \hat{\lambda}, \psi) = \sqrt{n}T(\mathbf{y}; \lambda, \psi) + B\sqrt{n}(\hat{\lambda} - \lambda) + o_p(1)$. Following Lemma 1, and assuming that a quantity bounded in probability has a finite expectation, we have $\mathrm{E}[\sqrt{n}T(\mathbf{y}; \hat{\lambda}, \psi)] = \mu_n + o(1)$ and $\mathrm{Var}[\sqrt{n}T(\mathbf{y}; \hat{\lambda}, \psi)] = V_n + BA_{11.2}^{-1}B' + o(1)$. These suggest that the statistic $\sqrt{n}T(\mathbf{y}; \hat{\lambda}, \psi)$ should be adjusted according to,

$$T^*(\mathbf{y}; \hat{\lambda}, \psi) = (V_n + BA_{11.2}^{-1}B')^{-\frac{1}{2}} \left( \sqrt{n}T(\mathbf{y}; \hat{\lambda}, \psi) - \mu_n + (V_n + BA_{11.2}^{-1}B')^{\frac{1}{2}} \mu_n \right),$$

which matches $V^{-\frac{1}{2}}\sqrt{n}T(\mathbf{y};\lambda,\psi)$ on its first and second moments with an error of $o(1)$. More importantly, the adjusted statistic should refer to the exact distribution of $\sqrt{n}T(\mathbf{y};\lambda,\psi)$ for an improved finite sample performance. Similar arguments apply to the case of constrained substitution.

## 4. Applications

In this section, we consider three applications to illustrate the generality and versatility of our theories. Some results turn out to be new and some are alternative derivations (using our theorems) of certain existing results.

### 4.1. Box-Cox Regression

The Box-Cox regression model (Box and Cox, 1964) is perhaps one of the models that best exemplify the applications of our methods, in particular Theorem 1. The usual Box-Cox transformation model has the form

$$h(\mathbf{y},\lambda) = \mathbf{X}\beta + \sigma\mathbf{e},$$

where $\mathbf{y}$ is an $n\times 1$ vector of original observations, $h(\mathbf{y},\lambda)$ is a vector of transformed observations, and $\mathbf{X}$ is an $n\times k$ matrix the columns of which contain the values of the explanatory variables $X_1, X_2, \ldots, X_k$, $\beta$ is a $k\times 1$ vector of regression coefficients, $\sigma$ is the error standard deviation, $\mathbf{e}$ is an $n\times 1$ vector of $N(0,1)$ variates, and $h(\cdot,\lambda)$ is a general monotonically increasing function, known except for $\lambda$. Note that the popular Box-Cox power transformation is incompatible with the model assumption as $e$ cannot be exactly normal unless $\lambda = 0$. For the choices of $h$ that do satisfy the model assumption, see Yeo and Johnson (2000).

It is clear that in this application, knowing the value of $\lambda$ greatly simplifies the inferences concerning $\beta$ or some functions of $\beta$ and $\sigma$. When $\lambda$ is known, exact inference methods are usually available. We demonstrate in this application how our theories can extend the $\lambda$-known inference methods to account for the estimation of

$\lambda$. The MLEs of $\beta$ and $\sigma^2$ for a given $\lambda$ are, respectively,

$$\hat{\beta}(\lambda) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'h(\mathbf{y}, \lambda) \quad \text{and} \quad \hat{\sigma}^2(\lambda) = n^{-1}\|\mathbf{M}h(\mathbf{y}, \lambda)\|^2,$$

where $\|\cdot\|$ is the Euclidian norm and $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ with $\mathbf{I}_n$ being the $n \times n$ identity matrix. The unconstrained MLE $\hat{\lambda}$ of $\lambda$ maximizes the concentrated log likelihood $L_{max}(\lambda) = -\frac{n}{2}\log\hat{\sigma}^2(\lambda) + \sum_{i=1}^n \log h_y(y_i, \lambda)$, where $h_y(y_i, \lambda) = \partial h(y_i, \lambda)/\partial y_i$. Likewise, the unconstrained MLEs of $\beta$ and $\sigma^2$ are, respectively, $\hat{\beta}(\hat{\lambda})$ and $\hat{\sigma}^2(\hat{\lambda})$. Let $\tilde{\sigma}^2(\lambda) = \frac{n}{n-k}\hat{\sigma}^2(\lambda)$ be the unbiased estimator of $\sigma^2$. Likewise, we have $\tilde{\sigma}^2(\hat{\lambda})$.

**Inferences concerning regression coefficients.** We consider the inferences for $\psi = a'\beta$, a general linear function of $\beta$ for a fixed vector $a$. When $\lambda$ is assumed known, we look to

$$\sqrt{n}\,T(\mathbf{y}; \lambda, \psi) = \frac{a'\hat{\beta}(\lambda) - \psi}{\{a'(\mathbf{X}'\mathbf{X})^{-1}a\}^{\frac{1}{2}}\tilde{\sigma}(\lambda)},$$

which is $t$ distributed with degrees of freedom $n-k$. The statistic $\sqrt{n}\,T(\mathbf{y}; \lambda, \psi)$ provides an exact $t$ test for testing $\psi$. It also leads to an exact and explicit CI for $\psi$. When $\lambda$ is unknown and is substituted by its unconstrained MLE $\hat{\lambda}$, we have

$$\sqrt{n}\,T(\mathbf{y}; \hat{\lambda}, \psi) = \frac{a'\hat{\beta}(\hat{\lambda}) - \psi}{\{a'(\mathbf{X}'\mathbf{X})^{-1}a\}^{\frac{1}{2}}\tilde{\sigma}(\hat{\lambda})}.$$

It is easy to verify that the conditions of Theorem 1 are satisfied. Hence, $\sqrt{n}\,T(\mathbf{y}; \hat{\lambda}, \psi)$ is asymptotically normal with mean zero and variance $1 + B^2 A_{11.2}^{-1}$, where

$$B = \lim_{n\to\infty} \frac{a'\mathrm{E}[\hat{\beta}_\lambda(\lambda)]}{\sqrt{n}\{a'(\mathbf{X}'\mathbf{X})^{-1}a\}^{\frac{1}{2}}\sigma},$$

and $\hat{\beta}_\lambda(\lambda)$ is the derivative of $\hat{\beta}(\lambda)$ with respect to $\lambda$. In practice, the above *variance inflation factor*, $B^2 A_{11.2}^{-1}$, can be easily estimated and $\sqrt{n}\,T(\mathbf{y}; \hat{\lambda}, \psi)$ can be corrected to have a $N(0, 1)$ limiting distribution, so that inference about $\psi$ based on the corrected statistic is asymptotically valid. In particular, a $100(1-\alpha)\%$ large sample CI for $\psi$ based on the corrected statistic $\sqrt{n}\,T(\mathbf{y}; \hat{\lambda}, \psi)(1 + \hat{B}^2\hat{A}_{11.2}^{-1})^{-\frac{1}{2}}$ takes the form

$$a'\hat{\beta}(\hat{\lambda}) \pm z_{\alpha/2}\{a'(\mathbf{X}'\mathbf{X})^{-1}a\}^{\frac{1}{2}}\tilde{\sigma}(\hat{\lambda})(1 + \hat{B}^2\hat{A}_{11.2}^{-1})^{\frac{1}{2}}. \tag{10}$$

Appeared in: Statistica Sinica, 2007, 17, 817-837.

Immediately following the arguments given in Section 3.4, and noting that the finite sample distribution of $\sqrt{n}\,T(\mathbf{y};\lambda,\psi)$ is an exact $t$ with $n-k$ degrees of freedom with $\mu_n = 0$ and $V_n = (n-k)/(n-k-2)$, one obtains an improved CI for $\psi$

$$a'\hat{\beta}(\hat{\lambda}) \pm t_{n-k}^{\alpha/2}\{a'(\mathbf{X}'\mathbf{X})^{-1}a\}^{\frac{1}{2}}\tilde{\sigma}(\hat{\lambda})(V_n + \hat{B}^2\hat{A}_{11.2}^{-1})^{\frac{1}{2}}. \tag{11}$$

However, use of the constrained estimator does not lead to this simple result even if the result of Theorem 2 is applicable. Bickel and Doksum (1981) showed that the asymptotic variance of $\hat{\beta}(\hat{\lambda})$ is larger than that of $\hat{\beta}(\lambda)$, and thus it is not valid for making inference concerning $\beta$ in the usual way. However, they did not provide ways to correct for the asymptotic variance of $\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\psi)$.

**Inference for the quantile function.** Suppose now we want to construct a confidence interval for $\psi = g(\lambda,\beta,\sigma^2) \equiv h^{-1}[(x_0'\beta + \sigma z_p),\lambda]$, the $p$-quantile of $y_0$ at a given observation $x_0$, where $z_p$ is the $p$-quantile of the standard normal variate. Note that $g$ now is a function of all the parameters. To state the problem in the framework of our theory, we need to find a statistic $T(\mathbf{y};\lambda,\psi)$ with a known distribution. A natural choice is

$$\sqrt{n}\,T(\mathbf{y};\lambda,\psi) = \frac{x_0'\hat{\beta}(\lambda) + \hat{\sigma}(\lambda)z_p - h(\psi,\lambda)}{\{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0\}^{\frac{1}{2}}\tilde{\sigma}(\lambda)},$$

which is distributed exactly as $t_{n-k}(-k_n z_p) + k_n z_p$, where $t_{n-k}(-k_n z_p)$ is a non-central $t$ with $n-k$ degrees of freedom and noncentrality parameter $-k_n z_p$, and $k_n = \{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0\}^{-\frac{1}{2}}$. Hence, when $\lambda$ is known an exact CI for $h(\psi,\lambda)$ can be constructed. Applying inverse transformations to the lower and upper confidence limits for $h(\psi,\lambda)$ gives the confidence limits for $\psi$. When $\lambda$ is unknown, substituting $\hat{\lambda}$ for $\lambda$ in the confidence limits results in a *plug-in* type of confidence interval. The validity of this interval depends on whether the statistic

$$\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\psi) = \frac{x_0'\hat{\beta}(\hat{\lambda}) + \hat{\sigma}(\hat{\lambda})z_p - h(\psi,\hat{\lambda})}{\{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0\}^{\frac{1}{2}}\tilde{\sigma}(\hat{\lambda})}$$

has the same limiting distribution as $\sqrt{n}\,T(\mathbf{y};\lambda,\psi)$. It can be verified that this problem fits into the framework of Theorem 1. Hence, $\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\psi)$ is asymptotically normal with mean zero and variance $1 + B^2 A_{11.2}^{-1}$, where

$$B = \lim_{n\to\infty} \frac{x_0'\mathrm{E}[\hat{\beta}_\lambda(\lambda)] + z_p\mathrm{E}[\hat{\sigma}_\lambda(\lambda)] - h_\lambda(\psi,\lambda)}{\sqrt{n}\{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0\}^{\frac{1}{2}}\sigma},$$

and $h_\lambda(\psi,\lambda)$ and $\hat{\sigma}_\lambda(\lambda)$ are the derivatives of $h(\psi,\lambda)$ and $\hat{\sigma}(\lambda)$ with respect to $\lambda$. Inference for $\psi$ can be made based on $\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\psi)/\sqrt{1+\hat{B}^2\hat{A}_{11.2}^{-1}}$ by simply referring the adjusted statistic to the standard normal distribution. Following the arguments given in Section 3.4, an improved inference for $\psi$ can be made based on

$$T^*(\mathbf{y};\hat{\lambda},\psi) = \frac{\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\psi) - \mu_n + \mu_n\sqrt{V_n + \hat{B}^2\hat{A}_{11.2}^{-1}}}{\sqrt{V_n + \hat{B}^2\hat{A}_{11.2}^{-1}}}$$

by referring $T^*(\mathbf{y};\hat{\lambda},\psi)$ to the distribution of $t_{n-k}(-k_n z_p) + k_n z_p$, which has a mean $\mu_n$ and a variance $V_n$ that can be obtained from the moments of a noncentral $t$ distribution. See Yang and Tse (2007) for the small and large sample results for a Box-Cox regression with heteroscedasticity. See also Yang and Tsui (2004) for a Box-Cox model in the context of modeling duration and event times.

It is clear from this application that it is much simpler to follow the unconstrained substitution approach to construct confidence interval for $\psi$ and that, even if one is concerned with hypothesis testing on $\psi$, the use of the constrained estimator $\hat{\lambda}_\psi$ as a replacement for $\lambda$ may have some disadvantages compared to using the unconstrained estimator $\hat{\lambda}$, as (i) the hypothesis imposed on $\psi$ does not simplify the estimation of $\lambda$, and (ii) a reparameterization has to be made to find the asymptotic variance of $\hat{\lambda}_\psi$, in particular in the case of quantiles.

Finally, as the concentrated log likelihood is available in this application, $A_{11.2}$ can be estimated using the simple method suggested in Section 3. Another interesting inference for this model may be the test of functional form, i.e., testing the value of $\lambda$, where the result of Theorem 2 may be applicable.

### 4.2. Dynamic Linear Regression with Serial Correlation

Dynamic linear regression with serial correlation is another example that illustrates the applications of our theories, simply because knowing the dynamic and serial correlation parameters reduces the model to a standard generalized least squares (GLS) regression. Also, this example can be used to illustrate the usefulness of Theorem 2, i.e., using constrained estimator does sometimes provide a simpler testing procedure than using unconstrained estimator. The model has the form

$$\begin{aligned} y_t &= \delta y_{t-1} + x_t'\beta + \varepsilon_t, \text{ with} \\ \varepsilon_t &= \rho\varepsilon_{t-1} + u_t, \ |\rho| < 1, \ t = 1, \cdots, n, \end{aligned}$$

where $x_t$ is a $k \times 1$ vector of independent variables and $\{u_t\}$ are a sequence of normal white noise with variance $\sigma^2$. For simplicity of exposition, we include only one lag in both $\{y_t\}$ and $\{\varepsilon_t\}$ processes. The results presented below are extendable to include more lags in both processes. Clearly in this application, knowing the values of $\delta$ and $\rho$ greatly simplifies the inferences concerning $\beta$. Also, the hypothesis $H_0$: $\rho = 0$ corresponds to an important test of model specification, under which estimation of $\delta$ and $\beta$ becomes much simpler.

**Inference concerning the regression coefficients.** Like the first application, we first consider the inference for $\psi = a'\beta$. In this context, we have $\lambda = (\delta, \rho)'$ and the other parameters besides $\psi$ and $\lambda$ are the nuisance parameters. Let $\mathbf{X}$ be the matrix of the fixed regressors. Define $y_t(\delta) = y_t - \delta y_{t-1}, t = 1, \cdots, n$, and let $\mathbf{y}(\delta) = \{y_t(\delta)\}_{n\times 1}$. Assume $y_0$ is fixed and $\{\varepsilon_1, \cdots, \varepsilon_n\}$ are stationary. We have $\mathbf{y}(\delta) \sim N(\mathbf{X}\beta, \sigma^2\Omega(\rho))$, where $\Omega(\rho)$ has elements $1/(1-\rho^2)$ in the diagonal and $\rho^{|i-j|}/(1-\rho^2)$ in the $(i,j)$ position. When $\lambda$ is known, the model reduces to a GLS regression. The constrained MLEs of $\beta$ (also the GLS) and $\sigma^2$ are given by

$$\begin{aligned} \hat\beta(\delta,\rho) &= (\mathbf{X}'\Omega^{-1}(\rho)\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}(\rho)\mathbf{y}(\delta), \\ \hat\sigma^2(\delta,\rho) &= n^{-1}[\mathbf{y}(\delta) - \mathbf{X}\hat\beta(\delta,\rho)]'\Omega^{-1}(\rho)[\mathbf{y}(\delta) - \mathbf{X}\hat\beta(\delta,\rho)]. \end{aligned}$$

The unconstrained MLEs of $\delta$ and $\rho$ are obtained by minimizing the concentrated log likelihood $L_{\max}(\delta,\rho) = -\frac{n}{2}\log\hat\sigma^2(\delta,\rho) - \frac{1}{2}\log|\Omega(\rho)|$. The unconstrained MLEs of $\beta$ and $\sigma^2$ are thus $\hat\beta(\hat\delta,\hat\rho)$ and $\hat\sigma^2(\hat\delta,\hat\rho)$, respectively. Also, $\hat\beta(\delta,\rho) \sim N[\beta, \sigma^2(\mathbf{X}'\Omega^{-1}(\rho)\mathbf{X})^{-1}]$,

Appeared in: Statistica Sinica, 2007, 17, 817-837.

which leads to an exact $t$ statistic for $\psi$:

$$\sqrt{n}T(\mathbf{y};\lambda,\psi) = \frac{a'\hat{\beta}(\delta,\rho)-\psi}{\{a'(\mathbf{X}'\Omega^{-1}(\rho)\mathbf{X})^{-1}a\}^{\frac{1}{2}}\hat{\sigma}(\delta,\rho)}.$$

When $\lambda$ is unknown and replaced by the unconstrained MLE to give the statistic $\sqrt{n}T(\mathbf{y};\hat{\lambda},\psi)$, we have from Theorem 1 that $\sqrt{n}T(\mathbf{y};\hat{\lambda},\psi)$ is asymptotically normally distributed with mean zero and variance $1+BA_{11.2}^{-1}B'$, where

$$B = \lim_{n\to\infty}\frac{a'\mathrm{E}[\partial\hat{\beta}(\delta,\rho)/\partial(\delta,\rho)]}{\sqrt{n}\{a'(\mathbf{X}'\Omega^{-1}(\rho)\mathbf{X})^{-1}a\}^{\frac{1}{2}}\sigma}.$$

Thus, if ordinary least squares standard errors are used for the transformed regression with consistently estimated $\delta$ and $\rho$, the standard errors are understated and the $t$-ratio is inflated (see Davidson and MacKinnon (1993, §10.4)). As the exact distribution of $\sqrt{n}T(\mathbf{y};\lambda,\psi)$ is known, the arguments given in Section 3.4 lead to the improved inference methods.

**Inference for serial correlation.** We now consider inference for the serial correlation parameter. We have the parameter of interest $\psi=\rho$, the parameters to be substituted $\lambda=(\delta,\beta')'$, and the nuisance parameter $\phi=\sigma^2$. To simplify the derivation, we assume (without loss of generality) that $\sigma=1$. We are interested in testing $H_0:\psi=0$. Consider the statistic

$$T(\mathbf{y};\lambda,\psi) = \frac{\sum_{t=1}^{n}\varepsilon_t\varepsilon_{t-1}}{\sum_{t=1}^{n}\varepsilon_t^2}-\psi, \tag{12}$$

where $\varepsilon_t=y_t-\delta y_{t-1}-x_t'\beta$. Note that $\sum_{t=1}^{n}\varepsilon_t\varepsilon_{t-1}/\sum_{t=1}^{n}\varepsilon_t^2$ is the constrained (on the index parameter $\lambda$, not $H_0$) MLE of $\psi$, and the conditions of Theorems 1 and 2 are satisfied. Under $H_0$, $T(\mathbf{y};\lambda,\psi)$ is asymptotically distributed as a standard normal variate. If $\mathbf{Z}=\{L\mathbf{y}:\mathbf{X}\}$ is the regression matrix including the lagged dependent vector $L\mathbf{y}$, then $A_{11}=\lim_{n\to\infty}\mathrm{E}[\mathbf{Z}'\mathbf{Z}/n]$. Let

$$\lim_{n\to\infty}\mathrm{E}[\mathbf{Z}'\mathbf{Z}/n] = \begin{pmatrix}\Sigma_{yy} & \Sigma_{yx}\\ \Sigma_{yx}' & \sigma_{xx}\end{pmatrix}.$$

It can be shown that $A_{21}=(1,0,...,0)$ and $A_{22}=1$. Furthermore, $B=(1,0,...,0)$ on $H_0$. Thus if we substitute the OLS estimate of $\lambda$, $\hat{\lambda}_\psi$ under $H_0:\psi=0$, into

Appeared in: Statistica Sinica, 2007, 17, 817-837.

$T(\mathbf{y}; \lambda, \psi)$, we conclude from Theorem 2 that $T(\mathbf{y}; \hat{\lambda}_\psi, \psi)$ is asymptotically normally distributed with mean 0 and variance $1 - v^2$, where

$$v^2 = BA_{11}^{-1}B' = 1/(\sigma_{yy} - \Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy}).$$

This result has been proved by Durbin (1970) in a more general context.

Now suppose we substitute the unconstrained MLE of $\lambda$ to obtain $T(\mathbf{y}; \hat{\lambda}, \psi)$. Then, from Theorem 1, on $H_0$ the asymptotic variance of $\sqrt{n}\,T(\mathbf{y}; \hat{\lambda}, \psi)$ is given by

$$1 + BA_{11.2}^{-1}B' = 1 + \frac{1}{(\sigma_{yy} - 1) - \Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy}}. \tag{13}$$

Note that $T(\mathbf{y}; \hat{\lambda}, \psi)$ is the unconstrained MLE of $\psi$, say $\hat{\psi}$. From standard MLE theory, the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi)$ is

$$(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} = \left(1 - \frac{1}{\sigma_{yy} - \Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy}}\right)^{-1},$$

which reduces to the expression in equation (10).

It is interesting to note that, contrary to inferences concerning the $\beta$ coefficients, the test for serial correlation based constrained substitution is simpler than that based on the unconstrained substitution, as the constrained estimator $\hat{\lambda}_\psi$ is much easier to calculate than the unconstrained estimator $\hat{\lambda}$. This phenomenon holds for many goodness-of-fit and residual-based diagnostic tests. The two tests are asymptotically equivalent under local alternatives, due to the asymptotic equivalence of the Lagrange multiplier and likelihood ratio tests. However, the estimated asymptotic variance of $T(\mathbf{y}; \hat{\lambda}_\psi, \psi)$ may be negative in small samples, especially when the exogenous variables are highly trended (see, for example, Tse (1985)). In contrast, the estimated asymptotic variance of $T(\mathbf{y}; \hat{\lambda}, \psi)$ is always positive.

The above results can be extended to cases where the residual variance is unknown and there are multiple lags in the dependent and error variables. While many model diagnostics are constructed based on the constrained MLE, mainly due to its simplicity in calculation, our results provide a way to obtain the asymptotic distribution of

Appeared in: Statistica Sinica, 2007, 17, 817-837.

a diagnostic when unconstrained MLE is used. In some cases, such as the tests for dynamic specification suggested by Sargan (1980), unconstrained MLE may be more convenient. Finally, our results may be applied to derive joint tests for the coefficients of the lags in the dependent and error variables.

### 4.3. Spatial Regression

Spatial econometrics has recently received much attention in the literature. Anselin (2001) gave a concise and informative survey. We use two commonly used spatial regression models: the *spatial lag model* and the *spatial error model*, to illustrate the applications of our theories. The spatial lag model takes the form

$$\mathbf{y} = \delta \mathbf{W}\mathbf{y} + \mathbf{X}\beta + \varepsilon, \tag{14}$$

where $\delta$ is the spatial autoregression coefficients, $\mathbf{W}$ is a given $n \times n$ matrix called the *spatial weights matrix*, and $\varepsilon \sim N(0, \sigma^2 I_n)$. $\mathbf{X}$ is the matrix of the regressors and $\beta$ is the vector of coefficients. Here $\lambda = \delta$ is the index parameter. Define $\mathbf{y}(\delta) = (I_n - \delta\mathbf{W})\mathbf{y}$. Then, when $\delta$ is given, the MLEs of $\beta$ and $\sigma^2$ have the same expressions as those in the Box-Cox regression, i.e., $\hat{\beta}(\delta) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}(\delta)$ and $\hat{\sigma}^2(\delta) = n^{-1}\|\mathbf{M}\mathbf{y}(\delta)\|^2$. The concentrated log likelihood is $L_{\max}(\delta) = -\frac{n}{2}\log\hat{\sigma}^2(\delta) + \log|I_n - \delta\mathbf{W}|$, which can be maximized to give the unconstrained MLE $\hat{\delta}$ of $\delta$. The unconstrained MLEs of $\beta$ and $\sigma^2$ are thus $\hat{\beta}(\hat{\delta})$ and $\hat{\sigma}^2(\hat{\delta})$, respectively. Inference for $\psi = a'\beta$ falls into the same framework as that of the Box-Cox regression. As before, it is inconvenient to apply Theorem 2 in this case to construct CI for $\psi$.

The spatial error model has the form

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \varepsilon, \\ \varepsilon &= \rho\mathbf{W}\varepsilon + \mathbf{u}, \end{aligned} \tag{15}$$

where $\rho$ is the error autoregressive coefficient and $\mathbf{u} \sim N(0, \sigma^2 I_n)$. Here $\lambda = \rho$ is the index parameter. It can be seen that $\mathrm{Var}(\varepsilon\varepsilon') = \sigma^2[(I_n - \rho\mathbf{W}')(I_n - \rho\mathbf{W})]^{-1} \equiv \sigma^2\Omega(\rho)$. So, when $\rho$ is known, the model (12) can be reduced to a linear regression model by pre-multiplying the matrix $I_n - \rho\mathbf{W}$ onto $\mathbf{y}$ and $\mathbf{X}$, and the constrained MLEs of $\beta$ and $\sigma^2$ are, respectively,

$$\begin{aligned} \hat{\beta}(\rho) &= [\mathbf{X}'(\rho)\mathbf{X}(\rho)]^{-1}\mathbf{X}'(\rho)\mathbf{y}(\rho), \text{ and} \\ \hat{\sigma}^2(\rho) &= n^{-1}[\mathbf{y}(\rho) - \mathbf{X}(\rho)\hat{\beta}(\rho)]'[\mathbf{y}(\rho) - \mathbf{X}(\rho)\hat{\beta}(\rho)], \end{aligned}$$

where $\mathbf{y}'(\rho) = \mathbf{y} - \rho\mathbf{W}\mathbf{y}$ and $\mathbf{X}'(\rho) = \mathbf{X} - \rho\mathbf{W}\mathbf{X}$ are the spatially filtered variables. Substituting $\hat{\beta}(\rho)$ and $\hat{\sigma}^2(\rho)$ into the log likelihood gives the concentrated log likelihood $L_{\max}(\rho) = -\frac{n}{2}\log\hat{\sigma}^2(\rho) + \sum_{i=1}^n \log(1 - \rho\omega_i)$, where $\omega_i$ are the eigenvalues of $W$. Maximizing $L_{\max}(\rho)$ gives the unconstrained MLE of $\rho$ which, upon substitution, gives the unconstrained MLEs of $\beta$ and $\sigma^2$ as $\hat{\beta}(\hat{\rho})$ and $\hat{\sigma}^2(\hat{\rho})$.

When inference concerns $\psi = a'\beta$, the exact $t$ statistic when $\rho$ is known takes the same form as that in the Box-Cox regression. Furthermore, it can be shown that $B = 0$. Hence, estimating $\rho$ (constrained or unconstrained) does not affect asymptotically the distribution of the test statistic. This is consistent with the fact that the information matrix for this model is block diagonal.

Other interesting inferences corresponding to the spatial regression model include i) testing $\delta = 0$ in the spatial lag model, ii) testing $\rho = 0$ in the spatial error model, and jointly testing for both $\delta = 0$ and $\rho = 0$ in a model where both types of spatial effects may exist. In these cases, the result of Theorem 2 may provide simpler solutions than does the result of Theorem 1.

Appeared in: Statistica Sinica, 2007, 17, 817-837.

## Appendix

**A.1 Derivation for Section 2** (Weibull Duration Model). In what follows, AVar denotes the asymptotic variance and ACov the asymptotic covariance. First, from the Taylor's expansion and the Law of Large Numbers, we have,

$$\begin{aligned}
\sqrt{n}\,T(\mathbf{y};\tilde{\lambda},\theta) &= \sqrt{n}\,T(\mathbf{y};\lambda,\theta)+\frac{1}{\lambda n}\sum_{i=1}^{n}\left[\left(\frac{y_i}{\theta}\right)^{\lambda}\log\left(\frac{y_i}{\theta}\right)^{\lambda}\right]\sqrt{n}(\tilde{\lambda}-\lambda)+o_p(1)\\
&= \sqrt{n}\,T(\mathbf{y};\lambda,\theta)+\frac{1}{\lambda n}\sum_{i=1}^{n}\mathrm{E}\left[\left(\frac{y_i}{\theta}\right)^{\lambda}\log\left(\frac{y_i}{\theta}\right)^{\lambda}\right]\sqrt{n}(\tilde{\lambda}-\lambda)+o_p(1)\\
&= \sqrt{n}\,T(\mathbf{y};\lambda,\theta)+\frac{1-\gamma}{\lambda}\sqrt{n}(\tilde{\lambda}-\lambda)+o_p(1)
\end{aligned}$$

for any $\sqrt{n}$-consistent estimator $\tilde{\lambda}$. The last equation follows from the result $\mathrm{E}[w\,\log w]=1-\gamma$, where $w$ is an exponential variable with mean 1 and $\gamma$ is Euler's constant.

Now the log-likelihood function is

$$L(\lambda,\theta)=n\log\lambda-n\lambda\log\theta+(\lambda-1)\sum_{i=1}^{n}\log y_i-\sum_{i=1}^{n}\left(\frac{y_i}{\theta}\right)^{\lambda},$$

which gives the score functions

$$\begin{aligned}
U_\lambda(\lambda,\theta) &= \frac{\partial L(\lambda,\theta)}{\partial\lambda}=\frac{n}{\lambda}+\sum_{i=1}^{n}\log\left(\frac{y_i}{\theta}\right)-\sum_{i=1}^{n}\left(\frac{y_i}{\theta}\right)^{\lambda}\log\left(\frac{y_i}{\theta}\right),\\
U_\theta(\lambda,\theta) &= \frac{\partial L(\lambda,\theta)}{\partial\theta}=-\frac{n\lambda}{\theta}+\frac{\lambda}{\theta}\sum_{i=1}^{n}\left(\frac{y_i}{\theta}\right)^{\lambda}.
\end{aligned}$$

The Fisher information matrix $I(\lambda,\theta)$ has the following elements: $I_{\lambda\lambda}=n[(1-\gamma)^2+(\pi^2/6)]/\lambda^2$, $I_{\lambda\theta}=I_{\theta\lambda}=-n(1-\gamma)/\theta$, and $I_{\theta\theta}=n(\lambda/\theta)^2$.

For (1), the unconstrained estimator $\hat{\lambda}$ involves both $U_\lambda$ and $U_\theta$. It can be easily seen to have the first-order approximation

$$\sqrt{n}\,(\hat{\lambda}-\lambda)=\sqrt{n}\,I^{\lambda\lambda}\left[U_\lambda(\lambda,\theta)+\frac{\theta(1-\gamma)}{\lambda^2}U_\theta(\lambda,\theta)\right]+o_p(1),$$

where $I^{\lambda\lambda}$ is the upper-left-corner block of $I^{-1}(\lambda,\theta)$. This, together with the fact that $T(\mathbf{y},\lambda,\theta)=\theta U_\theta(\lambda,\theta)/(n\lambda)$, leads immediately to $\mathrm{ACov}[T\,(\mathbf{y};\lambda,\theta),\sqrt{n}\,(\hat{\lambda}-\lambda)]=0$, and hence the result given in (1) with

$$\mathrm{AVar}[\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\theta)]=1+\left(\frac{1-\gamma}{\lambda}\right)^2\mathrm{AVar}[\sqrt{n}\,(\hat{\lambda}-\lambda)]=1+\frac{6(1-\gamma)^2}{\pi^2}.$$

Appeared in: Statistica Sinica, 2007, 17, 817-837.

For (2), the constrained estimator $\hat{\lambda}_\theta$ involves only $U_\lambda(\lambda,\theta)$ and hence has the first-order approximation

$$\sqrt{n}\,(\hat{\lambda}_\theta-\lambda)=\sqrt{n}\,I_{\lambda\lambda}^{-1}U_\lambda(\lambda,\theta)+o_p(1).$$

This gives, by noticing $T(\mathbf{y},\lambda,\theta)=\theta U_\theta(\lambda,\theta)/(n\lambda)$,

$$\mathrm{ACov}[\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\theta,\theta),\sqrt{n}\,(\hat{\lambda}_\theta-\lambda)]=\frac{\theta}{\lambda}I_{\lambda\lambda}^{-1}I_{\theta\lambda}+\frac{1-\gamma}{\lambda}nI_{\lambda\lambda}^{-1}=0.$$

Hence $\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\theta,\theta)$ is asymptotically independent of $\sqrt{n}\,(\hat{\lambda}_\theta-\lambda)$, which gives the result in (2) with

$$\mathrm{AVar}[\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\theta,\theta)]=1-\left(\frac{1-\gamma}{\lambda}\right)^2\mathrm{AVar}[\sqrt{n}\,(\hat{\lambda}_\theta-\lambda)]=1-\frac{(1-\gamma)^2}{(1-\gamma)^2+\pi^2/6}.$$

For the case of making inferences for $\psi$, we have, similar to the above,

$$\sqrt{n}\,T(\mathbf{y};\tilde{\lambda},\psi)=\sqrt{n}\,T(\mathbf{y};\lambda,\theta)+\left(\frac{1-\gamma}{\lambda}\frac{f_\lambda(\lambda,\psi)}{f(\lambda,\psi)}\right)\sqrt{n}(\tilde{\lambda}-\lambda)+o_p(1).$$

The result in (5) follows immediately from this expansion and the asymptotic independence between $\sqrt{n}\,T(\mathbf{y};\hat{\lambda},\theta)$ and $\sqrt{n}(\hat{\lambda}-\lambda)$ as shown above.

For the case of using $\hat{\lambda}_\psi$, using the results $U_\psi(\lambda,\psi)=U_\theta(\lambda,\theta)f_\theta(\lambda,\psi)$, $U_\lambda(\lambda,\psi)=U_\lambda(\lambda,\theta)+U_\theta(\lambda,\theta)f_\lambda(\lambda,\psi)$, $\sqrt{n}(\hat{\lambda}_\psi-\lambda)=\sqrt{n}{I^\circ}_{\lambda\lambda}^{-1}U_\lambda(\lambda,\psi)+o_p(1)$, and $T(\mathbf{y};\lambda,\psi)=T(\mathbf{y};\lambda,\theta)=\frac{\theta}{n\lambda}U_\theta(\lambda,\theta)$, one easily shows that $\mathrm{ACov}[\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\psi,\psi),\sqrt{n}(\hat{\lambda}_\psi-\lambda)]=0$. The result of (6) thus follows from ${I^\circ}_{\lambda\lambda}=I_{\lambda\lambda}+2f_\lambda I_{\theta\lambda}+f_\lambda^2 I_{\theta\theta}=n\left(\frac{\pi^2}{6\lambda^2}+(\frac{1-\gamma}{\lambda}-\frac{\lambda f_\lambda}{f})^2\right)$.

**A.2 Proof of Lemma 1:** First-order Taylor expansion on the joint likelihood equation $n^{-\frac{1}{2}}U(\hat{\lambda},\hat{\theta})=0$ leads to

$$\sqrt{n}\,(\hat{\lambda}-\lambda)=\frac{1}{\sqrt{n}}A_{11.2}^{-1}U_\lambda(\lambda,\theta)-\frac{1}{\sqrt{n}}A_{11.2}^{-1}A_{12}A_{22}^{-1}U_\theta(\lambda,\theta)+o_p(1),$$

It suffices to show that $\mathrm{AVar}[(\hat{\lambda}-\lambda)U_\theta(\lambda,\theta)']=0$, which follows directly from the asymptotic expansions given above:

$$\begin{aligned}\lim_{n\to\infty}\mathrm{E}[(\hat{\lambda}-\lambda)U_\theta(\lambda,\theta)'] &= \lim_{n\to\infty}\mathrm{E}[\frac{1}{n}A_{11.2}^{-1}(U_\lambda(\lambda,\theta)-A_{12}A_{22}^{-1}U_\theta(\lambda,\theta))U_\theta(\lambda,\theta)'] \\ &= A_{11.2}^{-1}(A_{12}-A_{12}A_{22}^{-1}A_{22})=0.\end{aligned}$$

Appeared in: Statistica Sinica, 2007, 17, 817-837.

**A.3 Proof of Theorem 1:** Assumption II and Lemma 1 lead immediately to the asymptotic independence of $\sqrt{n}T(\mathbf{y};\lambda,\psi)$ and $\sqrt{n}(\hat{\lambda}-\lambda)$, which together with the asymptotic normality of $\sqrt{n}T(\mathbf{y};\lambda,\psi)$ and $\sqrt{n}(\hat{\lambda}-\lambda)$ give the final result of Theorem 1.

**A.4 Proof of Theorem 2:** Assumption I gives

$$\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\psi,\psi) = \sqrt{n}\,T(\mathbf{y};\lambda,\psi) + B\sqrt{n}\,(\hat{\lambda}_\psi - \lambda) + o_p(1).$$

Similar to the expansion for $\hat{\lambda}$ given in the proof of Lemma 1, we have a first-order asymptotic expansion for $\hat{\lambda}_\psi$ in terms of the new parameterization:

$$\sqrt{n}\,(\hat{\lambda}_\psi - \lambda) = \frac{1}{\sqrt{n}} A^{\circ -1}_{11.2} U^\circ_\lambda(\lambda,\phi) - \frac{1}{\sqrt{n}} A^{\circ -1}_{11.2} A^\circ_{12} A^{\circ -1}_{22} U^\circ_\phi(\lambda,\phi) + o_p(1).$$

From the above we have

$$\begin{aligned} & \mathrm{ACov}[\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\psi,\psi), \sqrt{n}\,(\hat{\lambda}_\psi - \lambda)] \\ = \;& \mathrm{ACov}[\sqrt{n}\,T(\mathbf{y};\lambda,\psi), \sqrt{n}\,(\hat{\lambda}_\psi - \lambda)] + B\ \mathrm{AVar}[\sqrt{n}\,(\hat{\lambda}_\psi - \lambda)] \\ = \;& \mathrm{E}[\sqrt{n}\,T(\mathbf{y};\lambda,\psi)U^\circ_\lambda(\lambda,\phi)'] A^{\circ -1}_{11.2} - \mathrm{E}[\sqrt{n}\,T(\mathbf{y};\lambda,\psi)U^\circ_\phi(\lambda,\phi)'] A^{\circ -1}_{22} A^\circ_{21} A^{\circ -1}_{11.2} + BA^{\circ -1}_{11.2} \end{aligned}$$

Under Assumption IV, we apply the Dominated Convergence Theorem to obtain

$$\int \frac{\partial}{\partial \lambda'}(T(\mathbf{y};\lambda,\psi)p(\mathbf{y};\lambda,\psi,\phi))\,d\mathbf{y} = \frac{\partial}{\partial \lambda'}\int T(\mathbf{y};\lambda,\psi)p(\mathbf{y};\lambda,\psi,\phi)d\mathbf{y} = 0.$$

Thus, we have

$$\int \left(\frac{\partial}{\partial \lambda'}T(\mathbf{y};\lambda,\psi)\right) p(\mathbf{y};\lambda,\psi,\phi)d\mathbf{y} + \int T(\mathbf{y};\lambda,\psi)\left(\frac{\partial}{\partial \lambda'}p(\mathbf{y};\lambda,\psi,\phi)\right) d\mathbf{y} = 0.$$

As the second term on the RHS of the above equation is $\mathrm{E}[T(\mathbf{y};\lambda,\psi)U^\circ_\lambda(\lambda,\psi)]$, it follows that

$$\lim_{n\to\infty} \mathrm{E}[T(\mathbf{y};\lambda,\psi)U^\circ_\lambda(\lambda,\psi)'] = -\lim_{n\to\infty} \mathrm{E}[\partial T(\mathbf{y};\lambda,\psi)/\partial\lambda'] \equiv -B.$$

Similarly, we have $\lim_{n\to\infty}\mathrm{E}[T(\mathbf{y};\lambda,\psi)U^\circ_\phi(\lambda,\psi)'] = -\lim_{n\to\infty}\mathrm{E}[\partial T(\mathbf{y};\lambda,\psi)/\partial\phi'] = 0$, which shows $\mathrm{ACov}[\sqrt{n}\,T(\mathbf{y};\hat{\lambda}_\psi,\psi), \sqrt{n}\,(\hat{\lambda}_\psi - \lambda)] = 0$, completing the proof.

Appeared in: Statistica Sinica, 2007, 17, 817-837.

## Reference

Anselin, L. (2001). Spatial econometrics. In *A Companion to Theoretical Econometrics* (Baltagi, B. H. Eds.). Blackwell.

Bera, A. K. and Bilias, Y. (2001). Rao's score, Neyman's $C(\alpha)$ and Silvey's LM tests: an essay on historical developments and some new results. *Journal of Statistical Planning and Inference* **97**, 9-44.

Bera, A. K. and Kim, S. (2002). Testing constancy of correlation and other specifications of the BEARCH model with an application to international equity return. *Journal of Empirical Finance* **9**, 171-195.

Bera, A. K. and Zuo, X. (1996). Specification tests for a linear regression model with ARCH process. *Journal of Statistical Planning and Inference* **50**, 283-380.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformation revisited. *Journal of the American Statistical Association* **76**, 296-311.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformation (with discussion). *Journal of the Royal Statistical Society B* **26**, 211-252.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, New York.

Cox,D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics.* Oxford University Press, Oxford.

Durbin, J. (1970). Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrica* **38**, 410-421.

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics.* Cambridge University Press, Cambridge.

Kiefer, N. M. (1988). Economic Duration Data and Hazard Functions. *Journal of Economic Literature* **26**, 646-679.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics* **10**, 475-478.

Sargan, J. D. (1980). Some tests of dynamic specification for a single equation. *Econometrica* **48**, 879-897.

Tse, Y. K. (1985). Some modified versions of Durbin's $h$-statistic. *Review of Economics and Statistics* **67**, 534-538.

Tse, Y. K. (2002). Residual-based diagnostics for conditional heteroscedasticity models. *Econometrics Journal* **5**, 358-373.

Yang, Z. L. and Tse, Y. K. (2007). A corrected plug-in method for quantile interval construction through a transformed regression. *Journal of Business and Economic Statistics*, forthcoming.

Yang, Z. L. and Tsui, A. K. (2004). Analytically calibrated Box-Cox percentile limits for duration and event-time models. *Insurance Mathematics and Economics* **35**, 649-677.

Yeo, I. K. and Johnson, R. A. (2000). A new family of power transformation to improve normality or symmetry. *Biometrika* **87**, 954-959.