1990

Small Sample Properties of Estimators and Test Statistics in Nonlinear Regression: The Box-Cox Transformation.

Minbo Kim
Louisiana State University and Agricultural & Mechanical College

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313·761-4700 800·521-0600

Order Number 9123209

Small sample properties of estimators and test statistics in nonlinear regression: The Box-Cox transformation

Kim, Minbo, Ph.D.

The Louisiana State University and Agricultural and Mechanical Col., 1990

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

SMALL SAMPLE PROPERTIES OF ESTIMATORS AND TEST STATISTICS IN NONLINEAR REGRESSION: THE BOX-COX TRANSFORMATION

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirement for the degree of
Doctor of Philosophy

in

The Department of Economics

by
Minbo Kim
B.A., Seoul National University, 1980
M.S., Louisiana State University, 1988
December 1990

ACKNOWLEDGEMENTS

I wish to express my appreciation for the constructive comments supplied by Dr. Stephen W. Looney. I wish to thank Dr. David J. Smyth for summer grants and his faith in me.

I am especially grateful to my major professor R. Carter Hill for his encouragement, suggestions, and guidance of this dissertation and throughout my graduate study at Louisiana State University in Baton Rouge.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

The dissertation will address small sample properties of estimators and test statistics in a nonlinear regression model. The Box-Cox transformation is attractive to economists because a family of functional forms can be compared simultaneously within the framework of classical statistical inference. Usually, maximum likelihood (ML) methods are used to estimate the Box-Cox model. In the present study, nonlinear two stage and iterative generalized least squares (IGLS) method are considered. The accuracy of probability statements concerning nonlinear models is often questionable in small samples. Therefore, the finite sample distribution of the asymptotic t-statistic in the Box-Cox model is derived using an Edgeworth expansion. Bootstrapping, the more practical method for obtaining small sample distributions, is also discussed.

ML estimation of Box-Cox transformation suffers from a violation of the usual regularity conditions since the likelihood function of the Box-Cox model is not a proper density function. Since it is required that $y_t > 0$ in order for the Box-Cox transformation to be well-defined, the dependent variable is assumed to have a truncated normal distribution. The asymptotically equivalent covariance matrix estimators and test statistics— Lagrange multiplier, likelihood ratio and Wald —are compared in small samples.

The risk superiority of the Stein-rule estimator to the ML estimator is known in the context of the linear model. The usefulness of Stein-like estimation in the nonlinear Box-Cox model is investigated by considering the finite sample risk properties of ML and Stein-like estimators.

It is expected that this dissertation will make four major contributions to the current econometric literature. First, we introduce IGLS estimation of the Box-Cox model, and thus make linear statistical inference applicable to the nonlinear model. Second, the exact distribution of asymptotic t-ratios is derived and the bootstrap inversion of Edgeworth expansion is used. Third, the small sample distribution and power properties of three asymptotically equivalent test statistics are investigated. Fourth, shrinkage estimation is used in the determination of functional form.

CHAPTER 1

INTRODUCTION

Despair has no wings,
Nor has love,
No countenance:
They do not speak.
I do not stir,
I do not behold them,
I do not speak to them,
But I am as real as my love and my despair.

Paul Eluard

This research is concerned with the small sample properties of the Box-Cox transformation [Zarembka (1974); Poirier (1978); Spitzer(1978, 1982a)]. Since the statistical inference of nonlinear regression models relies on asymptotic theory, the accuracy of probability statements concerning nonlinear models is often questionable in small samples. There has been a continuous effort to develop an exact small sample distribution theory for estimators and test statistics. Basmann (1961) and Bergstrom (1962) initiated the interest in this area. Sargan (1975, 1976) and Phillips (1977) provided a theoretical foundation for approximate finite sample distributions using Edgeworth expansions. But the application of their results is limited since the approximations depend upon unknown parameters. Furthermore, the Edgeworth correction performs poorly when the errors in asymptotic results are large, while the correction works well in the region of the parameter space where the asymptotics provide a good approximation. The relevance of finite sample theory for practical econometrics is discussed by Taylor (1983). He casts doubt on the usefulness of the theoretical development of small sample distribu-

tions. Though a Monte Carlo experiment has difficulties in yielding general and precise results, the complexity of the problem suggests Monte Carlo simulation. In these regards, we wish to focus on an extensive Monte Carlo simulation to study the finite sample properties of the estimators and test statistics in the Box-Cox model.

In Chapter 2, the theoretical aspects of estimation and several estimation methods are surveyed. These results provide a basis for the remainder of the dissertation. The chapter includes discussions of point estimation, the linear regression model, maximum likelihood (ML) estimation, and asymptotic and small sample theory.

Chapter 3 is concerned with the estimation of the general transformation-of-variables model, and properties of its estimators and test statistics in small samples. Spitzer (1978) carried out a Monte Carlo simulation of the power transformation in the regression model. But his study is limited in the number of replications and the model specification. In the present study, the estimation methods are discussed in the context of the general power transformation model. The exact distribution of the t ratio will be derived analytically and investigated via Monte Carlo simulation, because it is widely used in applied econometric research.

Since it is required that $y_t > 0$ in order for the Box-Cox transformation to be well-defined, the density of y_t implied by the usual likelihood function is not proper. Therefore, the probability density of the original response (dependent)

variable needs to be properly defined to make the likelihood function regular and thus make the theoretical analysis of the Box-Cox transformation accurate. For our analysis of Box-Cox ML estimators, the response variable is defined to have a truncated normal distribution.

Recently, several investigators have studied the small sample properties of ML estimators obtained via alternative numerical optimization procedures. Griffiths, Hill and Pope (1987) examined the small sample properties of asymptotically equivalent covariance matrix estimators associated with different algorithms used to maximize the likelihood function [Newton-Raphson, method of scoring and the method of Berndt, Hall, Hall and Hausman (BHHH)] when the probit model is estimated. A similar study in the context of the simultaneous equation model was performed by Calzolari and Panattoni (1988). In general, their studies showed that the variance estimator obtained from BHHH is larger than that obtained from Newton-Raphson and reflects finite sample variability more accurately than does the estimated Newton-Raphson covariance, which is a better estimator of the true information matrix.

In Chapter 4, asymptotically equivalent covariance matrix estimators for the Box-Cox model will be studied in small samples. The Lagrange multiplier, likelihood ratio and Wald statistics are shown to be asymptotically equivalent under the null and local alternative hypothesis. Also, the empirical distribution of these asymptotic test statistics is investigated.

Under the squared error loss function, it is well known that the ML estimator is inadmissible for the case of more than two parameters in the linear regression model. Though the James-Stein estimator is a minimax estimator, it too is inadmissible. Baranchik (1964) proposed the positive part Stein-rule estimator, which has risk at least as small as the James-Stein estimator over the entire parameter space. Adkins and Hill (1989) investigated the risk properties of Stein-rule estimators for the probit model and found the same risk performance as for the linear regression model. But there are few other studies of Stein-like or other biased estimation procedures in the nonlinear regression model [Dagenais (1983); Schaefer et al (1984); Schaefer (1986)]. In Chapter 5, we compare full ML, constrained ML, pretest and Stein-like estimaton for the Box-Cox model in the context of risk properties. The asymptotic risks of these four estimators for the Box-Cox model will be derived analytically. The risk properties of these estimators will also be investigated in small samples.

Finally, we will give a summary and draw conclusions for this research in Chapter 6.

CHAPTER 2

STATISTICAL INFERENCE

In this chapter, we briefly summarize statistical inference concepts that are used in this dissertation.

2.1 Basic Definitions

The concept of a random experiment provides the starting point of a probability model. If we take the example of throwing a die, we can describe six possible outcomes, though the outcome of this experiment cannot be predicted *a priori*. When this kind of experiment can be repeated under identical conditions, we call it a random experiment.

Definition 2.1 *A set of all possible outcomes of a random experiment is defined to be the sample space* Ω. *A combination of elements of* Ω *is said to be an event.*

In probability models for uncountable spaces Ω, we need to assign probabilities to subsets of Ω, not to individual outcomes. In order to impose a mathematical structure on the set of events, the following definition is relevant:

Definition 2.2 *A system* $\mathcal{F}$ *of subsets of* Ω *is a* σ*-algebra, or* σ*-field, if it satisfies the following condition:*

1. $\Omega \in \mathcal{F}$
2. *If* $C_n \in \mathcal{F}$ $(n = 1, \cdots)$, *then* $\cup C_n \in \mathcal{F}$ $(or \cap C_n \in \mathcal{F})$

3. $C \in \mathcal{F} \Rightarrow \bar{C} \in \mathcal{F}$ *where* $\bar{C}$ *is a complement of* C.

Then the concept of probability is defined as follows:

Definition 2.3 *Probability* P *is defined as a probability measure (or set function) on* $\mathcal{F}$ *if*

1. $P(C) \geq 0$ *for any* $C \in \mathcal{F}$

2. $P(\Omega) = 1$

3. $P(\cup_{n=1}^{\infty} C_n) = \sum_{n=1}^{\infty} P(C_n)$ *for all mutually exclusive events in* $\mathcal{F}$.

Definition 2.4 *A probability model is an ordered triple* $(\Omega, \mathcal{F}, P)$ *where*

1. Ω *is a sample space*

2. $\mathcal{F}$ *is a* σ*-algebra of subsets of* Ω

3. P *is a probability measure on* $\mathcal{F}$

In addition, the following definitions are given for our analysis:

Definition 2.5 *Let* A *be any subset of* Ω. *The indicator (or characteristic) function of* A *is defined by*

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

The indicator function has the following properties:

1. $I_A(\omega) = 1 - I_{\bar{A}}(\omega)$ for any $A \in \mathcal{F}$ where $\bar{A}$ is the complement of A.

2. $I_{\cap_1^n A_i(\omega)} = \prod_{i=1}^{n} I_{A_i(\omega)}$ for $A_i \in \mathcal{F}$, $i = 1, \ldots, n$

3. $I_{\cup_1^n A_i(\omega)} = \max[I_{A_1}(\omega), \ldots, I_{A_n}(\omega)]$ for $A_i \in \mathcal{F}$, $i = 1, \ldots, n$

4. $I_A \cdot I_A = I_A$ for any $A \in \mathcal{F}$

Definition 2.6 *The collection of Borel sets $\mathcal{B}$ is the smallest σ-algebra which contains all of the open sets (or closed sets).*

Definition 2.7 *Let $\{I_n\}$ be the countable collections of open intervals that cover A such that*

$$A \subseteq \cup_{n=1}^{\infty} I_n$$

Then the outer measure of A is defined to be

$$m^*(A) = \inf_{A \subseteq \cup I_n} \sum \ell(I_n)$$

where $\ell(I_n)$ is the length of the interval I_n.

This definition implies that $m^*(\emptyset) = 0$ and that if $A \subseteq B$, then $m^*(A) \leq m^*(B)$. The set of a single point has outer measure zero, i.e. $m^*(\{r\}) = 0$ for any real number $r \in \Re$.

Definition 2.8 (Caratheódory) *A set M is said to be measurable if for each set S,*

$$m^*(S) = m^*(S \cap M) + m^*(S \cap \bar{M})$$

where $\bar{M}$ is the complement of M.

From this definition, it is straightforward to infer that $\bar{M}$ is measurable. The empty set and $\Re$ are also measurable.

Definition 2.9 *If M is a measurable set, the Lebesgue measure $m(M)$ is defined by the outer measure of M. If we restrict the set function m^* to the family of measurable sets, the set function m (Lebesgue measure) is obtained.*

Definition 2.10 *If the set of points where a property cannot hold has measure zero, then this property is said to hold almost everywhere (almost surely).*

Definition 2.11 *A function $\varphi = \varphi(\omega)$ defined on $(\Omega, \mathcal{F})$ is an $\mathcal{F}$-measurable function if*

$$\varphi^{-1}(X) = \{\omega : \varphi(\omega) \in X\}$$

is a measurable set in Ω.

Definition 2.12 *From definition* (2.11)*, if $(\Omega, \mathcal{F}) = (\Re, \mathcal{B}(\Re))$, then $\varphi(\omega)$ is a Borel function. Or equivalently, φ is said to be Borel measurable if for each $r \in \Re$ the set $\{\omega : \varphi(\omega) > r, r \in \Re\}$ is a Borel set.*

In the context of approximation, the notion of order of magnitude can be represented by O and o notation. These notations can be extended to the probabilistic approximation by use of O_p and o_p.

Definition 2.13 *For some positive constant C, the sequence $\{S_T\}$ is said to be of order $O(K_T)$ if*

$$\lim_{T \to \infty} \frac{|S_T|}{K_T} < C$$

Definition 2.14 *The sequence* $\{S_T\}$ *is said to be of order* $o(K_T)$ *if*

$$\lim_{T\to\infty} \frac{S_T}{K_T} = 0$$

Definition 2.15 *The sequence of random variables* $\{X_T\}$ *is said to be of order* $O_p(K_T)$ *if there exist a convergent nonstochastic sequence* $\{R_T\}$ *such that*

$$\frac{X_T}{K_T} - R_T \xrightarrow{p} 0$$

Definition 2.16 *The sequence of random variables* $\{X_T\}$ *is of order* $o_p(K_T)$ *if*

$$\frac{X_T}{K_T} \xrightarrow{p} 0$$

2.2 Point Estimation Theory

Suppose the independent and identically distributed random variables $X_1, \ldots, X_T$ have a common probability density which is of a fully specified parametric form and which depends on an unknown parameter θ. The family of probability density functions is then denoted by

$$\Xi = \{f(x;\theta),\ \theta \in \Theta\}, \quad \Theta \subseteq \Re$$

Point estimation is related to the problem of obtaining the best estimate of θ from the sample data $x_1, \ldots, x_n$. Therefore, point estimation has the form of a mapping $\varphi(\cdot) : \mathcal{A} \to \Theta$ where $\mathcal{A}$, is the set of real numbers $\{x : x = X(\omega), \omega \in \Omega\}$ and φ is a Borel function. An estimator is defined by $\varphi(X) : \Omega \to \Theta$ and its realized value $\varphi(x), x \in \mathcal{A}$, is called an estimate. Concerning the concept of the best estimator, we need to consider methods of constructing desirable estimators.

2.2.1 Finite Sample Properties

Suppose that a statistic or an estimator $\hat{\theta}$ for a scalar parameter θ has bias:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Then the mean squared error is defined by

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= V(\hat{\theta}) + B(\hat{\theta})^2 \end{aligned}$$

Definition 2.17 *An estimator $\hat{\theta}$ of θ is called an unbiased estimator if $E(\hat{\theta}) = \theta$.*

Under the restriction of unbiasedness, the best estimator can be defined as one which minimizes $V(\hat{\theta})$. But the minization of variance is not a completely relevant criterion since it is difficult to compare the dispersions of two distributions of different shapes in terms of variances.

Definition 2.18 *$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_T)$ is called a minimum variance unbiased estimator* (MVUE) *of θ if*

$$B(\hat{\theta}) = 0 \tag{2.1}$$

$$V(\hat{\theta}) \leq V(\bar{\theta}) \tag{2.2}$$

where $\bar{\theta}$ represents any other unbiased estimator of θ.

Consider the random sample X_1, X_2, X_3 from a normal distribution $N(\theta, 3)$, $\theta \in \Re$. The statistic $\bar{X} = \frac{X_1+X_2+X_3}{3}$ is distributed as $N(\theta, 1)$. Therefore, $\bar{X}$ is unbiased and $V(\bar{X}) \leq V(X_i)$, for i=1,2,3. But it is not certain that $\bar{X}$ is a minimum variance unbiased estimator. The Cramer-Rao inequality gives a basis for finding this type of estimator. Consider the following regularity conditions on Ξ:

1. The set $\{x : f(x;\theta) > 0\}$ does not depend on θ.

2. For each $\theta \in \Theta$, the derivatives $\frac{\partial^i \log f(\cdot)}{\partial \theta^i}$ (for $i = 1, 2, 3$) exist for all $x \in \mathcal{A}$.

3. Eq^2 exists for all $\theta \in \Theta$, i.e. $Eq^2 < \infty$, where $q = \frac{\partial \log f(\cdot)}{\partial \theta}$.

If $\hat{\theta}$ is an unbiased estimator, the Cramer-Rao inequality has the form:

$$V(\hat{\theta}) \geq [Eq^2]^{-1}$$

where $[Eq^2]^{-1}$ is the Cramer-Rao lower bound. If we consider the previous example again, $q = \frac{\sum_{i=1}^{3}(X_i-\theta)}{3}$ and thus $Eq^2 = V(\bar{X}) = 1$. Since $\bar{X}$ satisfies the Cramer-Rao lower bound, we know that $\bar{X}$ is a minimum variance unbiased estimator. Furthermore, the restriction of linearity of estimators can reduce the size of the class of estimators. In our example, $\bar{X}$ is a linear combination of random variables; therefore, this estimator is called a best linear unbiased estimator. In general, the Cramer-Rao bound is provided as

$$CR(\hat{\theta}) = [1 + \frac{dB(\hat{\theta})}{d\theta}]^2 / Eq^2$$

Definition 2.19 *Let X and Y be arbitrary random variables of finite variance. Suppose that $EX = m_1$ and $EY = m_2$. Then the linear function $\alpha + \beta X$ which predicts Y has minimum mean square error $E(Y - \alpha - \beta X)^2$ when $\alpha = m_2 - \frac{COV(X,Y)}{V(X)} m_1$ and $\beta = \frac{COV(X,Y)}{V(X)}$. The equation*

$$\hat{Y} = m_2 + \frac{COV(X,Y)}{V(X)}(X - m_1)$$

is called the second-order regression equation.

We can write Y using the second-order regression equation

$$Y = \hat{Y} + Z \tag{2.3}$$

where

$$\begin{aligned} E(Z) &= 0 \\ COV(\hat{Y}, Z) &= 0 \\ V(Z) &= V(Y) - \frac{[COV(X,Y)]^2}{V(X)} \end{aligned}$$

Therefore Y is a linear function of X if and only if

$$V(Y) = \frac{[COV(X,Y)]^2}{V(X)}$$

If Y is not a linear function of X, the Cauchy-Schwarz inequality is

$$V(Y) > \frac{[COV(X,Y)]^2}{V(X)}$$

since $V(Z) > 0$.

Proposition 2.1 *$V(\hat{\theta}) \geq CR(\hat{\theta})$ for any estimator $\hat{\theta}$, where equality holds if $\hat{\theta}$ and $q(\theta)$ are linear functions of each other* [Cox and Hinkley (1974)].

Proof:

We consider

$$E(\hat{\theta}) = \int \hat{\theta} f(x;\theta)dx = \theta + B(\hat{\theta})$$

Differentiating both sides with respect to θ, *we get*

$$\begin{aligned} 1 + \frac{dB(\hat{\theta})}{d\theta} &= \int \hat{\theta} \frac{df(x;\theta)}{d\theta} \frac{1}{f(x;\theta)} f(x;\theta)dx \\ &= \int \hat{\theta} \frac{d \log f(x;\theta)}{d\theta} f(x;\theta)dx \\ &= COV(\hat{\theta}, q(\theta)) \end{aligned}$$

By use of Cauchy-Schwarz *inequality,*

$$[COV(\hat{\theta}, q(\theta))]^2 \leq V(\hat{\theta})V(q(\theta))$$

Therefore,

$$\frac{(1 + \frac{dB(\hat{\theta})}{d\theta})^2}{V(\hat{\theta})E(q^2)} \leq 1$$

since $V(q) = E(q^2)$.

Next, we discuss the case where the Cramer-Rao lower bound of unbiased estimators is attainable.

Definition 2.20 *Let the statistic* $S = s(X)$ *and a family of* $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ *be the family of possible distributions of* X *where* $X = (X_1, \ldots, X_T)$. *Then a statistic* S *is said to be sufficient for* X, *or for the family* $\mathcal{P}$, *if the conditional distribution of* X *given* $S = s(x)$ *is independent of* θ.

Let S be a sufficient statistic and T, which is not a function of S alone, be any unbiased estimator for θ. Let $E(T|s) = \varphi(s)$. Then the conditional expectation $Y = \varphi(S)$ is a function of only S. The sufficiency of S is required lest Y should depend on the parameter θ. According to the Rao-Blackwell theorem, for any given θ,

$$\begin{aligned} E(Y;\theta) &= E(T;\theta) = \theta \\ V(Y;\theta) &\leq V(T;\theta) \end{aligned}$$

where equality holds if and only if T is a function of S almost surely.

Definition 2.21 *Consider a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{G}$ be the collection of all subsets G of Ω such that $F_1 \subseteq G \subseteq F_2$ and $P(F_2 \setminus F_1) = 0$ for any $F_1, F_2 \in \mathcal{F}$. Then a probability measure P is complete if $\mathcal{G} = \mathcal{F}$.*

For example, let the random variable X have a density function that is contained in the family of probability densities $\{h(x;\theta);\ \theta \in \Theta\}$ where $h(\cdot)$ is a Borel measurable function. The family of probability densities $\{h(x;\theta);\ \theta \in \Theta\}$ is said to be a complete family if $E[w(x)] = 0$ implies that $w(x) = 0$ almost surely for every $\theta \in \Theta$.

Suppose two different unbiased estimators T_1 and T_2 yield $\varphi_1(s) = E(T_1|S = s)$ and $\varphi_2(s) = E(T_2|S = s)$. Then $Y_1 = \varphi_1(S)$ and $Y_2 = \varphi_2(S)$ are functions of S such that

$$E(Y_1 - Y_2;\theta) = 0 \quad \forall\theta$$

If S has a complete density function, $Y_1 = Y_2$ almost surely. Therefore, we can deduce that any function of a sufficient statistic with a complete density is the unique minimum unbiased estimator of its expectation.

In practice, a minimum variance unbiased estimator can be easily identified within the exponential class of probability density functions.

Definition 2.22 *Consider a probability model* $(\Omega, \mathcal{F}, P)$. *Let* $K_1, \ldots, K_k$ *be* k *measurable functions over* Ω. *Then the probability density function*

$$f(\omega, \theta) = C(\theta)h(\omega)\exp[\sum_{i=1}^{k} \pi_i(\theta)K_i(\omega)], \quad \omega \in \Omega \text{ and } \theta \in \Theta,$$

where the $\pi_i(\theta)$ *are given nontrivial functions, is said to be a member of the exponential class of probability density functions with respect to the probability measure* P.

We consider a family of one parameter continuous density functions:

$$\begin{array}{rcll} f(x;\theta) & = & \exp[a(\theta)K(x) + R(x) + b(\theta)], & x \in (c,d) \\ & = & 0, & \textit{otherwise} \end{array}$$

where $\theta \in \Theta \subseteq \Re$. Suppose that the regularity conditions of this exponential class are given by

1. c and d are not a function of θ.

2. $a(\theta)$ is a continuous function of θ and the case when $a(\theta) = 0$ is excluded.

3. $\frac{d}{dx}K(\cdot) \neq 0$ and $R(x)$ is continuous.

Let $X_1, \ldots, X_T$ denote a random sample from a distribution which is a member of the regular exponential class of continuous density functions. Then the joint probability density of $X_1, \ldots, X_T$ can be written as

$$\exp[a(\theta)\sum_{i=1}^{T} K_i(x) + \sum_{i=1}^{T} R_i(x) + Tb(\theta)]$$

By the factorization theorem [Hogg and Craig (1978), p. 344], $S = \sum_{i=1}^{T} K_i(x)$ is a sufficient statistic for θ and the probability density of S is easily shown to be complete. Therefore, if we can find an unbiased form of $\varphi(s)$, then the unique MVUE of θ can be identified. Let us take the previous example of three random variables X_1, X_2, X_3, each distributed as $N(\theta, 3)$. The joint density of X_1, X_2, X_3 is

$$\begin{aligned} f(x_1, x_2, x_3; \theta) &= (\frac{1}{\sqrt{2\pi \cdot 3}})^3 \exp[-\frac{\sum_{i=1}^{3}(x_i - \theta)^2}{2 \cdot 3}] \\ &= \exp[\frac{\theta}{3}\sum_{i=1}^{3} x_i - \frac{1}{6}\sum_{i=1}^{3} x_i^2 - \frac{3}{2}\ln(6\pi) - \frac{\theta^2}{2}] \end{aligned}$$

Thus, $\bar{X} = \frac{\sum_{i=1}^{3} X_i}{3}$ is the unique MVUE for θ since $S = \sum_{i=1}^{3} X_i$ is a complete and sufficient statistic for θ and $E(S/3) = \theta$.

Suppose the random variables $X_1, \ldots, X_T$ from a common distribution F have realizations $x_1, \ldots, x_T$. Let $Y = \hat{\theta}_T(X_1, \ldots, X_T)$ be an estimator for θ in Θ and let $\delta(y)$ denote the function which assigns an action for a given $y = \hat{\theta}_T(x_1, \ldots, x_T)$. Then $\delta(Y)$ is called a decision function or decision rule and a specific value $\delta(y)$ is a decision. In the context of the estimation problem, it would be useful to have a measure of the difference between a parameter θ and the decision rule (point

estimator) $\delta(Y)$. Let $\mathcal{L}(\theta, \delta(y))$ be a nonnegative loss function. The risk function is given by

$$\rho(\theta, \delta) = E_F(\mathcal{L}[\theta, \delta(Y)]) = \int_{-\infty}^{+\infty} \mathcal{L}[\theta, \delta(y)] f(y; \theta) dy$$

if Y is a continuous random variable.

Definition 2.23 *An estimator $\hat{\theta}_0$ is said to dominate an estimator $\hat{\theta}_1$ if*

$$\rho(\theta, \hat{\theta}_0) \leq \rho(\theta, \hat{\theta}_1), \quad \forall \theta \in \Theta$$

Furthermore, $\hat{\theta}_0$ strictly dominates $\hat{\theta}_1$ if strict inequality holds for some $\theta \in \Theta$.

Usually, it is impossible to choose a decision rule which minimizes the risk function for every value of θ in Θ.

Definition 2.24 *An estimator is said to be admissible if it is not strictly dominated by any other estimator of θ.*

If we adhere to the principle of choosing a decision rule which minimizes the maximum risk, a useful estimator can be defined:

Definition 2.25 *An estimator $\hat{\theta}_0$ is said to be minimax if*

$$\sup_{\theta \in \Theta} \rho(\theta, \hat{\theta}_0) \leq \sup_{\theta \in \Theta} \rho(\theta, \hat{\theta})$$

for any other estimator $\hat{\theta}$ of θ.

For example, a minimax estimator under the condition that $E(\hat{\theta}_0) = \theta$ and $\mathcal{L}(\theta, \delta(y)) = (\theta - \delta(y))^2$ gives a minimum variance unbiased estimator. In particular, the decision rule $\delta(Y)$ which minimizes $E[\theta - \delta(Y)]^2$ for all θ in Θ is called the minimum mean squared error estimator. In the context of minimum mean squared error estimation, an estimator of small bias and small variance is preferred to the unbiased estimator having considerable variance.

2.2.2 Jackknife and Bootstrapping

The jackknife method is based on the idea that the study of the change in an estimator resulting from the elimination of observations enables us to judge the stability of the estimator. Suppose $X_1, \ldots, X_T$ are independent and identically distributed random variables from an unknown probability distribution F on some space. Let $\hat{\theta}_T$ be an estimator for a parameter θ obtained from observations $X_1 = x_1, \ldots, X_T = x_T$, i.e. $\hat{\theta}_T = \hat{\theta}_T(x_1, \ldots, x_T)$. Then a deleted estimator can be obtained by deleting each x_i sequentially:

$$\hat{\theta}_T^{(i)} = \hat{\theta}_T(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_T)$$

Let

$$\hat{\theta}_T^{(\cdot)} = \frac{1}{T} \sum_{i=1}^{T} \hat{\theta}_T^{(i)}$$

The estimator of bias of $\hat{\theta}_T$ is

$$\hat{B}(\hat{\theta}_T) = (T-1)[\hat{\theta}_T^{(\cdot)} - \hat{\theta}_T]$$

The bias-corrected jackknife estimator is written as

$$\hat{\theta}_T^J = \hat{\theta}_T - \hat{B}(\hat{\theta}_T) = T\hat{\theta}_T - (T-1)\hat{\theta}_T^{(\cdot)}$$

It is well known that many of the usual statistics (e.g., the ML estimator) have

$$\begin{aligned} E(\hat{\theta}_T) &= \theta + \frac{a_1}{T} + \frac{a_2}{T} + \cdots \\ &= \theta + O(T^{-1}) \end{aligned} \tag{2.4}$$

where the a_i do not depend on T [Schucany, Gray and Owen (1971)]. Thus,

$$\begin{aligned} E(\hat{\theta}_T^J) &= TE(\hat{\theta}_T) - (T-1)E(\hat{\theta}_T^{(\cdot)}) \\ &= \theta - \frac{a_2}{T(T-1)} + \cdots \\ &= \theta + O(T^{-2}) \end{aligned} \tag{2.5}$$

From Equations (2.4) and (2.5), we see that the jackknife estimator has bias of order T^{-2} while the original estimator is biased up to T^{-1}. The estimated jackknife variance suggested by Tukey (1958) is given by

$$\hat{V}(\hat{\theta}_T) = \frac{T-1}{T}\sum_{i=1}^{T}(\hat{\theta}_T^{(i)} - \hat{\theta}_T^{(\cdot)})^2 \tag{2.6}$$

Example:

1. Assume that $\mathcal{A} = \Re$, $\theta = E(X) = \int_{\Re} X dF$. Let $\hat{\theta} = \bar{X}$ where $\bar{X} = \sum X_i/T$. Since $\hat{\theta}_T^{(\cdot)} = \bar{X}$, $\hat{\theta}_T^J = \bar{X}$.

2. Assume $\theta = \int_{\Re}(X - EX)^2 dF$. Let $\hat{\theta}_T = \frac{\sum(X_i - \bar{X})^2}{T}$. Since $\hat{B}(\hat{\theta}_T) = -\frac{\sum(X_i - \bar{X})^2}{T(T-1)}$, $\hat{\theta}_T^J = \frac{\sum(X_i - \bar{X})^2}{T}$.

Now consider the classical regression model

$$y_t = \underline{x}_t'\underline{\beta} + e_t$$

where $e_t \sim iid\ F$, $E_F e_t = 0$ and $\underline{x}_t$ is a $K \times 1$ vector of regressors at the tth observation. The OLS estimators are

$$\begin{aligned} \underline{\hat{\beta}} &= (\sum_{t=1}^{T} \underline{x}_t\underline{x}_t')^{-1} \sum_{t=1}^{T} \underline{x}_t y_t \\ \hat{V}(\underline{\hat{\beta}}) &= \sum_{t=1}^{T} \frac{\hat{e}_t^2}{T-K} (\sum_{t=1}^{T} \underline{x}_t\underline{x}_t')^{-1} \end{aligned}$$

where $\hat{e}_t = y_t - \underline{x}_t'\underline{\hat{\beta}}$. Let $\underline{\hat{\beta}}^{(t)}$ be an OLS estimator obtained from $(T-1)$ observations by deleting the tth observation:

$$\underline{\hat{\beta}}^{(t)} = \underline{\hat{\beta}} - (\sum_{t=1}^{T} \underline{x}_t\underline{x}_t')^{-1}\underline{x}_t \frac{\hat{e}_t}{1 - \underline{x}_t'(\sum_{t=1}^{T} \underline{x}_t\underline{x}_t')^{-1}\underline{x}_t}$$

The jackknife covariance estimator of $\underline{\hat{\beta}}$ is given by

$$(T-1)/T \sum_{t=1}^{T} [\underline{\hat{\beta}}^{(t)} - \underline{\hat{\beta}}^{(\cdot)}][\underline{\hat{\beta}}^{(t)} - \underline{\hat{\beta}}^{(\cdot)}]'$$

where $\underline{\hat{\beta}}^{(\cdot)} = \frac{1}{T}\sum_{t=1}^{T} \underline{\hat{\beta}}^{(t)}$. This can be modified as

$$\frac{T-1}{T}(X'X)^{-1}[X'D^*X - \frac{1}{T}(X'\underline{\hat{e}}^*\underline{\hat{e}}^{*\prime}X)](X'X)^{-1} \tag{2.7}$$

where

$$\begin{aligned} X &= (\underline{x}_1, \ldots, \underline{x}_T)' \\ D^* &= diag(\hat{e}_1^{*^2}, \ldots, \hat{e}_T^{*^2}) \\ \underline{\hat{e}}^* &= (\hat{e}_1^*, \ldots, \hat{e}_T^*)' \\ \hat{e}_t^* &= \hat{e}_t/[1 - \underline{x}_t'(\sum_{t=1}^{T} \underline{x}_t\underline{x}_t')^{-1}\underline{x}_t] \end{aligned}$$

MacKinnon and White (1985) showed via Monte Carlo simulation that in the presence of heteroscedastic error disturbances, the performance of the t ratio based on the jackknife variance estimator (2.7) was good in small samples. Furthermore, these t ratios were robust to the heteroscedasticity relative to those based on the other heteroscedasticity-consistent covariance estimators. Kleijnen et al (1987) jackknifed the feasible generalized least squares estimator and obtained better confidence intervals.

Bootstrapping generalizes the idea of the deleted statistic $\hat{\theta}_T^{(i)}$. By using the empirical distribution $\hat{F}$, which is assumed to have probability mass $1/T$ at each data point x_i, we can draw a random sample with replacement from the observations $x_1, \ldots, x_T$ such that $x_1^*, \ldots, x_T^*$ are independent and identically distributed according to $\hat{F}$. Then the estimator $\hat{\theta}_T$ for a parameter θ can be calculated many (B) times. These simulated statistics $\hat{\theta}_T^{*1}, \ldots, \hat{\theta}_T^{*B}$ are used to estimate the distribution F. The bootstrap expected value of $\hat{\theta}_T$ is

$$E_*(\hat{\theta}_T) = \sum_{b=1}^{\mathrm{B}} \hat{\theta}_T^{*b}/\mathrm{B}$$

Therefore, we have the bootstrap bias of $\hat{\theta}_T$

$$B_*(\hat{\theta}_T) = E_*(\hat{\theta}_T) - \hat{\theta}_T$$

and the bootstrap variance and standard deviation are

$$\begin{aligned} V_*(\hat{\theta}_T) &= \sum_{b=1}^{\mathrm{B}} (\hat{\theta}_T^{*b} - E_*(\hat{\theta}_T))^2/(\mathrm{B}-1) \qquad (2.8) \\ SD_*(\hat{\theta}_T) &= \sqrt{V_*(\hat{\theta}_T)} \end{aligned}$$

The nonparametric ML estimator [Efron (1981a)] of the standard deviation $\sigma(\hat{F})$ is equivalent to $SD_*(\hat{\theta}_T)$ as $T \to \infty$. Bootstrap theory can also be developed in a parametric framework such as

$$\hat{F}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi(\sum(x_i - \bar{x})^2/T)}} \exp(-\frac{(z-\bar{x})^2}{2\sum(x_i - \bar{x})^2/T})dz$$

In the context of a regression model, Freedman (1981) showed that bootstrapping the linear regression model gave the same asymptotics as the classical least squares estimators. He also showed that the bootstrap produced the desired asymptotic result even in the presence of heteroscedasticity in the disturbances when the model contained endogenous regressors and when instrumental variables were introduced. Beran (1982) obtained the first order Edgeworth expansion of a bootstrap estimator and showed that it was asymptotically equivalent to the usual bootstrap estimator. His study provided the theoretical evidence of the validity of bootstrapping in small samples.

Freedman and Peters (1984a, 1984b) carried out bootstrap simulation to check the bias of standard errors from feasible generalized least squares (FGLS), two-stage least squares (2SLS) and three-stage least squares (3SLS) methods in small samples. Standard errors of FGLS and 2SLS were very small relative to the bootstrap standard deviation while those of 3SLS did not indicate any substantial differences from the bootstrap finite sample variability. The bootstrap method was also applied in testing the linear restriction in the context of a seemingly unrelated regression model [Rocke (1989)]. Rocke's experimental investigation bore

witness to the good performance of the bootstrap Bartlett adjustment method in small samples. He used Rothenberg's second-order asymptotic method for the Bartlett adjustment [Bartlett (1937); Cox (1984); Rothenberg (1984a)].

2.3 Estimators in the Linear Regression Model

2.3.1 Ordinary Least Squares (OLS)

The classical linear regression model is written as

$$\underline{y} = X\underline{\beta} + \underline{\epsilon} \tag{2.9}$$

under the following assumptions:

1. X is a nonstochastic matrix of full column rank, which is less than the number of observations T.

2. There exists a nonsingular matrix $Q = \lim_{T\to\infty} \frac{1}{T}(X'X)$

3. The vector $\underline{\epsilon}$ consists of random disturbances satisfying $E(\underline{\epsilon}) = \underline{0}$ and $V(\underline{\epsilon}) = \sigma^2 I$.

The least squares estimator $\hat{\underline{\beta}}$ is the value of $\underline{\beta}$ which minimizes the Euclidian distance between $\underline{y}$ and $X\underline{\beta}$:

$$D(\underline{y}, X\underline{\beta}) = (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})$$

Thus the analytical solution yields

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$$

where $E\underline{\hat{\beta}} = \underline{\beta}$ and $V(\underline{\hat{\beta}}) = \sigma^2(X'X)^{-1}$.

Theorem 2.1 (Gauss-Markov) *Let C be a $K \times T$ nonstochastic matrix satisfying $CX = I$. Define $\underline{\tilde{\beta}} = C\underline{y}$. Then $E(\underline{\tilde{\beta}} - \underline{\beta})(\underline{\tilde{\beta}} - \underline{\beta})' - E(\underline{\hat{\beta}} - \underline{\beta})(\underline{\hat{\beta}} - \underline{\beta})'$ is positive definite if $\underline{\tilde{\beta}} \neq \underline{\hat{\beta}}$. $\underline{\hat{\beta}}$ is called the best linear unbiased estimator* (BLUE).

Proof:

$$\begin{aligned} E\underline{\tilde{\beta}} &= E(CX\underline{\beta} + C\underline{\epsilon}) = \underline{\beta}, \quad \textit{for any } C \\ E(\underline{\tilde{\beta}} - \beta)(\underline{\tilde{\beta}} - \beta)' &= E(C\underline{\epsilon}\underline{\epsilon}'C') \\ &= \sigma^2 CC' \\ &= \sigma^2(X'X)^{-1} + \sigma^2(C - (X'X)^{-1}X')(C - (X'X)^{-1}X')' \\ &= E(\underline{\hat{\beta}} - \underline{\beta})(\underline{\hat{\beta}} - \underline{\beta})' + \sigma^2[(C - (X'X)^{-1}X')(C - (X'X)^{-1}X')'] \end{aligned}$$

If $\underline{\epsilon}$ is assumed to have normal distribution $N(\underline{0}, \sigma^2 I)$, then the least squares estimator is the ML estimator and the covariance matrix of the ML estimator attains a Cramer-Rao lower bound.

2.3.2 Generalized Least Squares (GLS)

The assumption of scalar covariance random disturbances is too strict to be realistic. Therefore, we will consider more general assumptions about the error disturbances:

1. $\underline{\epsilon}$ is a $T \times 1$ vector of random disturbances with $E\underline{\epsilon} = \underline{0}$ and $V(\underline{\epsilon}) = \Sigma$, where Σ is a known $T \times T$ symmetric positive definite matrix.

2. There exists a finite nonsingular matrix $Q = \lim_{T \to \infty} \frac{1}{T}(X'\Sigma^{-1}X)$.

Theorem 2.2 (Aitken) *The estimator defined as*

$$\tilde{\underline{\beta}} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\underline{y}$$

is efficient within the class of linear unbiased estimators of $\underline{\beta}$.

Proof:

Since Σ *is positive definite,* $\Sigma^{-1/2}$ *can be defined as* $P\Lambda^{-1/2}P'$ *where* P *is an orthogonal matrix of characteristic vectors of* Σ, *and* Λ *is a diagonal matrix composed of the eigenvalues of* Σ. *Transform the original model* (2.9) *by premultiplying by* $\Sigma^{-1/2}$ *on both sides:*

$$\begin{aligned} \Sigma^{-1/2}\underline{y} &= \Sigma^{-1/2}X\underline{\beta} + \Sigma^{-1/2}\underline{\epsilon} \\ \underline{y}^* &= X^*\underline{\beta} + \underline{\epsilon}^* \end{aligned} \tag{2.10}$$

where $E\underline{\epsilon}^* = \underline{0}$ *and* $V(\underline{\epsilon}^*) = I$. *Then the usual least squares estimation can be applied to the transformed model* (2.10):

$$\begin{aligned} \tilde{\underline{\beta}} &= (X^{*'}X^*)^{-1}X^{*'}\underline{y}^* \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\underline{y} \end{aligned}$$

where $E(\tilde{\underline{\beta}}) = \underline{0}$ *and* $V(\tilde{\underline{\beta}}) = (X'\Sigma^{-1}X)^{-1}$. *The Gauss-Markov theorem holds for the transformed model.*

In practice, the estimated covariance matrix is used for estimation since the covariance matrix of the GLS model is unknown. The FGLS estimator is

$$\underline{\tilde{\tilde{\beta}}} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}\underline{y}$$

where $V(\underline{\tilde{\tilde{\beta}}}) = (X'\hat{\Sigma}^{-1}X)^{-1}$ and $\hat{\Sigma}$ is a consistent estimator for Σ. When the covariance matrix of the GLS estimator is unknown, the Aitken theorem does not hold. In small samples, the standard errors from FGLS were shown to be too optimistic (small) when compared to the small sample variability [Freedman and Peters (1984a)].

2.3.3 Biased Estimation

Traditionally, unbiased least squares estimators are preferred to biased estimators. However, if the statistical model is incorrectly specified, then the least squares estimator may be a biased estimator. Ignorance of non-sample information, therefore, may yield suboptimal estimators. Stein (1961) proposed a biased estimator which dominates the least squares estimator under quadratic loss. It is a common practice to use pretest estimators in linear regression models when *a priori* information is considered. When *a priori* information is tested and accepted using an appropriate test statisic, the restricted least squares estimator is used; otherwise the least squares estimator is used. But the pretest estimator does not dominate the least squares estimator in a risk context. We next discuss the pretest and Stein-rule estimators.

Pretest Estimator

The classical regression model (2.9) can be reparametrized as

$$\begin{aligned} \underline{y} &= XS^{-1/2}S^{1/2}\underline{\beta} + \underline{e} \\ &= Z\underline{\theta} + \underline{e} \end{aligned} \tag{2.11}$$

where S represents a positive definite matrix $X'X$, $Z = XS^{-1/2}$ and $\underline{\theta} = S^{1/2}\underline{\beta}$.

Definition 2.26 *The weighted risk function under squared error loss is defined by*

$$\rho(\hat{\underline{\theta}}; \underline{\theta}, \sigma^2, Q) = E(\hat{\underline{\theta}} - \underline{\theta})'Q(\hat{\underline{\theta}} - \underline{\theta})/\sigma^2$$

where Q is any known positive definite matrix. If $Q = I$, then the unweighted risk function is defined by

$$\rho(\hat{\underline{\theta}}; \underline{\theta}, \sigma^2) = E(\hat{\underline{\theta}} - \underline{\theta})'(\hat{\underline{\theta}} - \underline{\theta})/\sigma^2$$

and the least squares estimator $\hat{\underline{\theta}} = Z'\underline{y} = S^{1/2}\hat{\underline{\beta}}$ has risk

$$\rho(\hat{\underline{\theta}}; \underline{\theta}, \sigma^2) = tr[(Z'Z)^{-1}] = tr(I_K) = K$$

Consider the test statistic u for testing the null hypothesis $H_0 : \underline{\theta} = \underline{r}$ against $H_1 : \underline{\theta} \neq \underline{r}$. Then $u \sim F_{(K,T-K)}$ under the null hypothesis. If H_0 is not true, u has a noncentral F distribution with noncentrality parameter $\lambda = (\underline{\theta} - \underline{r})'(\underline{\theta} - \underline{r})/(2\sigma^2)$. Let c be determined by $\int_c^\infty dF_{(K,T-K)} = \alpha$ where α is a significance level. Then the pretest estimator has the form:

$$\hat{\underline{\theta}} = \begin{cases} \underline{r} & \text{if } u < c \\ \hat{\underline{\theta}} & \text{if } u \geq c \end{cases}$$

or

$$\hat{\underline{\theta}} = I_{(0,c)}(u)\underline{r} + I_{[c,\infty)}(u)\hat{\underline{\theta}} = \hat{\underline{\theta}} - I_{(0,c)}(u)(\hat{\underline{\theta}} - \underline{r})$$

where $u = \frac{(T-K)(\hat{\underline{\theta}}-\underline{r})'(\hat{\underline{\theta}}-\underline{r})}{K(\underline{y}-Z\underline{\theta})'(\underline{y}-Z\hat{\underline{\theta}})}$ and $\hat{\underline{\theta}}$ is an unrestricted least squares estimator.

Proposition 2.2 *If the* $(K \times 1)$ *random vector* $\underline{w}$ *is distributed as multivariate normal with mean* $\underline{\delta}$ *and covariance matrix* $\sigma^2 I_K$ *and is independent of the random variable with the* $\chi^2_{(T-K)}$*-distribution, then*

$$E[I_{(0,c)}(\frac{\underline{w}'\underline{w}}{\sigma^2\chi^2_{(T-K)}})\frac{T-K}{K}\frac{\underline{w}}{\sigma}] = \frac{\underline{\delta}}{\sigma}P[\chi^2_{(K+2,\lambda)}/\chi^2_{(T-K)} < c\frac{K}{T-K}]$$

where $\lambda = \frac{\underline{\delta}'\underline{\delta}}{2\sigma^2}$ *and* $\underline{\delta} = \underline{\theta} - \underline{r}$.

Proof:

Use the following result from Judge and Bock (1978, pp. 321-322)*:*

$$E[\varphi(\frac{\underline{w}'\underline{w}}{\sigma^2})(\frac{\underline{w}}{\sigma})] = (\frac{\underline{\delta}}{\sigma})E[\varphi(\chi^2_{(K+2,\lambda)})]$$

where $\varphi(\cdot)$ *is a Borel measurable function. In addition, note that*

$$E[I_{(-\infty,c)}(u)] = P[u < c]$$

Then the expected value of the pretest estimator can be obtained using this proposition:

$$E(\hat{\underline{\theta}}) = \underline{\theta} - (\underline{\theta} - \underline{r})P[\chi^2_{(K+2,\lambda)}/\chi^2_{(T-K)} < cK/(T-K)]$$

Therefore, the pretest estimator is biased unless $\underline{\delta} = \underline{0}$ or $P = 0$. Since $\underline{\delta}$ is the bias of the restricted least squares estimator and P is not greater than 1, the bias

of the pretest estimator is less than or equal to that of the restricted least squares estimator under the exact linear restriction: $\underline{\theta} = \underline{r}$. The risk function of the pretest estimator can be written as [Judge and Bock (1978), p. 71]

$$\begin{aligned}\rho(\hat{\underline{\theta}}; \underline{\theta}, \sigma^2) &= K + (2\frac{\underline{\delta}'\underline{\delta}}{\sigma^2} - K)P[\chi^2_{(K+2,\lambda)}/\chi^2_{(T-K)} \leq cK/(T-K)] \\ &\quad - (\frac{\underline{\delta}'\underline{\delta}}{\sigma^2})P[\chi^2_{(K+4,\lambda)}/\chi^2_{(T-K)} \leq cK/(T-K)] \\ &= K + (2\frac{\underline{\delta}'\underline{\delta}}{\sigma^2} - K)h_\lambda(2) - (\frac{\underline{\delta}'\underline{\delta}}{\sigma^2})h_\lambda(4)\end{aligned}$$

where $h_\lambda(\ell) = P[\chi^2_{(K+\ell,\lambda)}/\chi^2_{(T-K)} \leq cK/(T-K)]$ and $0 < h_\lambda(4) < h_\lambda(2) < 1$. Therefore, if the null hypothesis $H_0 : \underline{\delta} = \underline{0}$ is correct, the risk of the pretest estimator is less than that of the least squares estimator. Let the hypothesis error be defined as the distance of the hypothesized value from the true parameter value. As hypothesis error $\underline{\delta}'\underline{\delta}$ increases, the risk of the pretest estimator becomes increasingly higher than that of the least squares estimator and approaches the risk of the least squares estimator with an infinite error of hypothesis. There is no one dominant estimator among the least squares, restricted least squares and pretest estimators under quadratic error loss measured over the entire range of hypothesis error.

Stein-rule Estimator

For convenience of analysis, premultiply both sides of equation (2.11) by Z':

$$\begin{aligned}Z'\underline{y} &= Z'Z\underline{\theta} + Z'\underline{e} \\ \underline{z} &= \underline{\theta} + \underline{v} \qquad (2.12)\end{aligned}$$

where $\underline{v} \sim N(\underline{0}, \sigma^2 I)$. Then the ML estimator $\hat{\underline{\theta}} = Z'\underline{y} = \underline{z}$ and the risk of the ML estimator is

$$\rho(\hat{\underline{\theta}}; \underline{\theta}, \sigma^2) = K$$

Suppose σ^2 is known to be 1. Then the Stein-rule estimator has the form:

$$\ddot{\underline{\theta}} = [1 - a/(\hat{\underline{\theta}}'\hat{\underline{\theta}})]\hat{\underline{\theta}}$$

When σ^2 is unknown, the analogous Stein-rule estimator is

$$\ddot{\underline{\theta}} = [1 - as/(\hat{\underline{\theta}}'\hat{\underline{\theta}})]\hat{\underline{\theta}}$$

where $s = \underline{y}'[I - X(X'X)^{-1}X']\underline{y}$ and a is a constant taking a value in the range of $0 \leq a \leq 2(K-2)/(T-K+2)$ for minimaxity. The risk for this estimator is

$$\begin{aligned}
\rho(\ddot{\underline{\theta}}; \underline{\theta}, \sigma^2) &= E\frac{(\hat{\underline{\theta}} - \underline{\theta})'(\hat{\underline{\theta}} - \underline{\theta})}{\sigma^2} - 2aE[\frac{s}{\sigma^2}] + 2a\underline{\theta}'E[\frac{s}{\sigma^2}\hat{\underline{\theta}}/(\hat{\underline{\theta}}'\hat{\underline{\theta}})] \\
&\quad + a^2E[\frac{s^2}{\sigma^2}/(\hat{\underline{\theta}}'\hat{\underline{\theta}})] \\
&= K + a(T-K)(T-K+2)[a - \frac{2(K-2)}{(T-K+2)}]E[1/\chi^2_{(K,\lambda)}]
\end{aligned}$$

where $\lambda = \underline{\theta}'\underline{\theta}/2\sigma^2$. If $K \geq 3$ and $0 \leq a \leq \frac{2(K-2)}{(T-K+2)}$, then $\rho(\ddot{\underline{\theta}}; \underline{\theta}, \sigma^2) \leq \rho(\hat{\underline{\theta}}; \underline{\theta}, \sigma^2)$. Therefore, the ML estimator is inadmissible. The minimum risk of the Stein-rule estimator is attained at $a = \frac{K-2}{T-K+2}$ with the corresponding risk of

$$\rho(\ddot{\underline{\theta}}; \underline{\theta}, \sigma^2) = K - (K-2)^2(T-K)/(T-K+2)E[1/\chi^2_{(K,\lambda)}]$$

Note that $\hat{\underline{\theta}}'\hat{\underline{\theta}}/\sigma^2$ is distributed as $\chi^2_{(K,\lambda)}$. The optimal Stein-rule estimator becomes

$$\ddot{\underline{\theta}}_0 = [1 - \frac{(T-K)(K-2)}{(T-K+2)K}(\frac{1}{u})]\hat{\underline{\theta}}$$

where $u = \frac{(T-K)\hat{\underline{\theta}}'\hat{\underline{\theta}}}{Ks} \sim F_{(K,T-K,\lambda)}$. In this form, the Stein-rule estimator uses the test statistic u to combine the observed sample information with the *a priori* restriction $\underline{\theta} = \underline{0}$. This estimator differs from the pretest estimator in that the Stein-rule uses the test statistic u to shrink the unrestricted estimator toward the restricted estimator, while the pretest estimator chooses between the restricted and unrestricted estimator, depending on the test statistic. In general, if the shrinkage point is given by $\underline{\theta}_1$, the Stein-rule estimator has the form:

$$\ddot{\underline{\theta}}_1 = [1 - \frac{(K-2)(T-K)}{K(T-K+2)}(\frac{1}{u_1})](\hat{\underline{\theta}} - \underline{\theta}_1) + \underline{\theta}_1$$

where

$$u_1 = (T-K)(\hat{\underline{\theta}} - \underline{\theta}_1)'(\hat{\underline{\theta}} - \underline{\theta}_1)/Ks$$

Though the Stein-rule estimator is a minimax estimator, it is not admissible under squared error loss since other estimators have risk at least as small as the Stein-rule estimator [Baranchik (1964)]. Baranchik (1964) suggested a positive-rule estimator which improves on the Stein-rule estimator:

$$\ddot{\underline{\theta}}^+ = I_{[c,\infty)}(u)[1 - c/u]\hat{\underline{\theta}}$$

where $u = \frac{(T-K)\hat{\underline{\theta}}'\hat{\underline{\theta}}}{Ks}$ and $\frac{(K-2)(T-K)}{K(T-K+2)} < c \leq \frac{2(K-2)(T-K)}{K(T-K+2)}$. Now the risk of a positive part Stein-rule estimator is

$$\rho(\ddot{\underline{\theta}}^+; \underline{\theta}, \sigma^2) = \rho(\ddot{\underline{\theta}}; \underline{\theta}, \sigma^2) - 2(\frac{\underline{\theta}'\underline{\theta}}{\sigma^2})E[I_{(0,c)}(\frac{\chi^2_{(K+2,\lambda)}}{\chi^2_{(T-K)}})(c\frac{\chi^2_{(T-K)}}{\chi^2_{(K+2,\lambda)}} - 1)]$$

$$- E[I_{(0,c)}(\frac{\chi^2_{(K,\lambda)}}{\chi^2_{(T-K)}})(1 - c\frac{\chi^2_{(T-K)}}{\chi^2_{(K,\lambda)}})^2\chi^2_{(K,\lambda)}]$$

This positive part estimator can be extended to the general shrinkage estimator.

$$\ddot{\underline{\theta}}_1^+ = I_{[c,\infty)}(u_1)(1 - \frac{c}{u_1})(\hat{\underline{\theta}} - \underline{\theta}_1) + \underline{\theta}_1$$

where $u_1 = (T-K)(\hat{\underline{\theta}} - \underline{\theta}_1)'(\hat{\underline{\theta}} - \underline{\theta}_1)/Ks$.

Consider the original linear statistical model

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

In the context of the classical linear regression model, the traditional Stein-rule estimator is defined by

$$\underline{\delta}_1 = [1 - \frac{a(T-K)}{u_1 K}]\underline{b}$$

where

$$\begin{aligned} u_1 &= (T-K)\underline{b}'S\underline{b}/Ks \\ b &= S^{-1}X'\underline{y} \\ S &= X'X \\ s &= \underline{y}'(I - X(X'X)^{-1}X')\underline{y} \\ 0 \leq\ a &\leq \frac{[trS^{-1} - 2\eta_{max}(S^{-1})]}{(T-K+2)\eta_{max}(S^{-1})} \\ \eta_{max}(S^{-1}) &= \text{the largest characteristic root of } S^{-1} \end{aligned}$$

We introduce the linear restriction

$$R\underline{\beta} = \underline{r}$$

where R is a $J \times K$ matrix of rank J. Then the restricted least squares (RLS) estimator is

$$\underline{\hat{\beta}}_R = \underline{b} - S^{-1}R'(RS^{-1}R')^{-1}(R\underline{b} - \underline{r})$$

and the general minimax estimator which shrinks towards the RLS estimator is

$$\underline{\delta}_3 = [1 - \frac{a(T-K)}{u_3 J}](\underline{b} - \underline{\hat{\beta}}_R) + \underline{\hat{\beta}}_R$$

where $u_3 = \frac{(T-K)(R\underline{b}-\underline{r})'(RS^{-1}R')^{-1}(R\underline{b}-\underline{r})}{Js} \sim F_{(J,T-K,\lambda)}$ and $\lambda = \frac{(R\underline{\beta}-\underline{r})'(RS^{-1}R')^{-1}(R\underline{\beta}-\underline{r})}{2\sigma^2}$.

When $R = I_K$,

$$\underline{\delta}_2 = [1 - \frac{a(T-K)}{u_2 K}](\underline{b} - \underline{r}) + \underline{r}$$

where $u_2 = \frac{(T-K)(\underline{b}-\underline{r})'S(\underline{b}-\underline{r})}{Ks}$. Under the unweighted risk function, the general Stein-rule estimators ($\underline{\delta}_2$ and $\underline{\delta}_3$) are minimax and have risk improvement over the ML estimator $\underline{b}$ if $J \geq 3$ and $0 \leq a \leq \frac{2}{(T-K+2)}[\frac{trA}{\eta_{max}(A)} - 2]$ with $A = (RS^{-1}R')^{-1}RS^{-1}QS^{-1}R'$. The positive part general Stein-rule estimator is written as

$$\underline{\delta}^+ = I_{[a,\infty)}(u_3^*)[1 - \frac{a}{u_3^*}](\underline{b} - \underline{\hat{\beta}}_R) + \underline{\hat{\beta}}_R$$

where $u_3^* = \frac{J}{T-K}u_3$.

In general, Stein-rule estimators improve the risk over the ML estimator when we assume that the disturbances are normally distributed. Brandwein and Strawderman (1978, 1980) showed that risk improvement over the OLS estimator could be obtained in a class of spherically symmetrical disturbance distributions. Ullah et al (1983) studied the different conditions under which the Stein-rule estima-

tor dominates the OLS estimators in the case of nonnormal disturbances with moments up to the fourth order.

Neither the asymptotic nor exact distributions of the Stein-rule estimator are sufficiently well known to be used as a basis for statistical test procedures. Ullah (1974) derived the exact moments of the Stein-rule estimator $\underline{\delta}^* = (1 - \frac{as}{\underline{\hat{\theta}}'\underline{\hat{\theta}}})\underline{\hat{\theta}}$ to the fourth order. His finding was that if the original parameter θ_i is positive, then δ_i^* is negatively skewed and vice versa; if θ_i is close to zero, δ_i^* has a normal distribution. Phillips (1984) obtained the exact density of the Stein-rule estimator $\underline{\delta}^*$ under the condition of normality of error terms. Knight (1986) extended Phillips' analysis by assuming nonnormal disturbances which are well approximated by Edgeworth or Gram-Charlier distributions. However, these studies have no practical implications for hypothesis testing or confidence interval construction.

2.4 Maximum Likelihood Estimation and Asymptotic Properties

2.4.1 Asymptotic Properties

Definition 2.27 *An estimator $\hat{\theta} = h(X_1, \ldots, X_T)$ is said to be (weakly) consistent for θ if $\hat{\theta}$ converges in probability to the parameter θ, i.e.*

$$\lim_{T \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$$

Definition 2.28 *An estimator $\hat{\theta}$ is said to be strongly consistent if $\hat{\theta}$ converges*

almost surely to the parameter θ, i.e.

$$P(\lim_{T\to\infty} \hat{\theta} = \theta) = 1$$

If a ML function attains a global maximum, then the corresponding ML estimator has the consistency property.

Proposition 2.3 *Assume that:*

1. *The parameter space is a compact subset of $\Re^K$ and the true parameter $\underline{\theta}_0$ is in Θ.*
2. *The log-likelihood function $\ell(\underline{\theta}, \underline{y}, X)$ is continuous in $\underline{\theta} \in \Theta$ for all $\underline{y}$ and is a Borel measurable function of $\underline{y}$ $\forall \underline{\theta} \in \Theta$.*
3. *$\ell(\underline{\theta})/T$ converges in probability to a nonstochastic function $\ell_\infty(\underline{\theta})$ uniformly in $\theta \in \Theta$ as $T \to \infty$, and $\ell_\infty(\underline{\theta})$ has a unique global maximum at $\underline{\theta}_0$.*

Then the ML estimator $\hat{\underline{\theta}}$ converges in probability to $\underline{\theta}_0$.

Proof: [Amemiya (1985)]

Let $N(\underline{\theta}_0)$ be an open neighborhood of $\underline{\theta}_0$ in $\Re^K$. Then $\ell_m = \max_{\underline{\theta} \in \overline{N} \cap \Theta} \ell_\infty(\underline{\theta})$ exists since $\overline{N} \cap \Theta$ is a compact set where $\overline{N}$ is the complement of $N(\underline{\theta}_0)$. Let $\epsilon = \ell_\infty(\underline{\theta}_0) - \ell_m$. If $|\frac{1}{T}\ell - \ell_\infty| < \epsilon/2$ for all $\underline{\theta}$, then

$$\ell_\infty(\hat{\underline{\theta}}_T) > \frac{1}{T}\ell(\hat{\underline{\theta}}_T) - \epsilon/2 \tag{2.13}$$

and

$$\frac{1}{T}\ell(\underline{\theta}_0) > \ell_\infty(\underline{\theta}_0) - \epsilon/2 \tag{2.14}$$

Since $\ell(\underline{\theta})$ *attains a global maximum at* $\hat{\underline{\theta}}$, $\ell(\hat{\underline{\theta}}) \geq \ell(\underline{\theta}_0)$. *Let* $E_T = \{|\frac{1}{T}\ell(\underline{\theta}) - \ell_\infty(\underline{\theta})| < \epsilon/2, \forall\underline{\theta}\}$. Equation (2.13) *can be modified as:*

$$\ell_\infty(\hat{\underline{\theta}}_T) > \frac{1}{T}\ell(\underline{\theta}_0) - \epsilon/2 \tag{2.15}$$

From (2.14) *and* (2.15), E_T *implies that*

$$\ell_\infty(\hat{\underline{\theta}}) > \ell_\infty(\underline{\theta}_0) - \epsilon$$

Therefore,

$$\ell_\infty(\hat{\underline{\theta}}) > \ell_m$$

Consequently,

$$\hat{\underline{\theta}} \in N(\underline{\theta}_0)$$

Since $E_T \subseteq \{\hat{\underline{\theta}} \in N(\underline{\theta}_0)\}$,

$$P\{|\frac{1}{T}\ell(\underline{\theta}) - \ell_\infty(\underline{\theta})| < \epsilon/2, \forall\theta\} \quad \leq \quad P\{\hat{\underline{\theta}} \in N(\underline{\theta}_0)\}$$

From the assumption, $\lim_{T\to\infty} P(E_T) = 1$. *Consequently,* $\hat{\underline{\theta}}$ *converges in probability to* $\underline{\theta}_0$.

In practice, it is difficult to obtain a global ML estimator and to ascertain the global maximum using currently available optimization techniques. Therefore, we wish to prove the consistency and asymptotic normality of a local ML estimator.

Proposition 2.4 *Let Θ_T be the set of local maxima satisfying the first order condition for the likelihood function. Then under appropriate assumptions, for any $\epsilon > 0$,*

$$\hat{\underline{\theta}} \xrightarrow{p} \underline{\theta}_0 \text{ uniformly in } \Theta.$$

In addition,

$$\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) \rightarrow N(\underline{0}, H(\underline{\theta}_0)^{-1} J(\underline{\theta}_0) H(\underline{\theta}_0)^{-1})$$

where $H = -\lim_{T\to\infty} E\frac{1}{T}\frac{\partial^2 \ell}{\partial\underline{\theta}\partial\underline{\theta}'}|_{\underline{\theta}_0}$ *and* $J = \lim_{T\to\infty} E\frac{1}{T}\frac{\partial \ell}{\partial\underline{\theta}}\frac{\partial \ell}{\partial\underline{\theta}'}|_{\underline{\theta}_0}$.

Assumptions:

1. *Consider any parameter space $\Theta \subseteq \Re^K$ which contains the true parameter $\underline{\theta}_0$. Θ is assumed to be an open set.*

2. *The log-likelihood function ℓ is Borel measurable and $\ell \in C^1(N(\underline{\theta}_0))$ where C^1 denotes the class of functions which are continuous in their first derivatives and N denotes an open neighborhood.*

3. *$(\frac{\partial^2 \ell}{\partial\underline{\theta}\partial\underline{\theta}'})$ exists and is continuous in an open and convex neighborhood of $\underline{\theta}_0$.*

4. *$\frac{1}{T}\frac{\partial^2 \ell}{\partial\underline{\theta}\partial\underline{\theta}'}|_{\hat{\underline{\theta}}}$ converges in probability to a finite nonsingular negative definite matrix $\lim_{T\to\infty} E\frac{1}{T}\frac{\partial^2 \ell}{\partial\underline{\theta}\partial\underline{\theta}'}|_{\underline{\theta}_0}$ for any consistent estimator $\hat{\underline{\theta}}$.*

5. *$\frac{1}{T}\frac{\partial \ell}{\partial\underline{\theta}} \rightarrow N(\underline{0}, J(\underline{\theta}_0))$.*

6. *$\frac{1}{T}\ell(\underline{\theta})$ converges in probability to the nonstochastic function ℓ_∞ uniformly in $\underline{\theta} \in N(\underline{\theta}_0)$.*

7. $\text{plim}\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'}$ *exists and is continuous in* $N(\underline{\theta}_0)$.

Proof:

Consistency:

By a series expansion around $\underline{\theta}_0$, *we have*

$$\frac{1}{T}\ell(\underline{\theta}) = \frac{1}{T}\ell(\underline{\theta}_0) + \frac{1}{T}\frac{\partial \ell}{\partial \underline{\theta}'}\Big|_{\underline{\theta}_0}(\underline{\theta} - \underline{\theta}_0) + \frac{1}{2}(\underline{\theta} - \underline{\theta}_0)'\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'}\Big|_{\underline{\theta}^*}(\underline{\theta} - \underline{\theta}_0)$$

where $\|\underline{\theta}^* - \underline{\theta}_0\| \leq \|\underline{\theta} - \underline{\theta}_0\|$. *Therefore,*

$$\text{plim}\frac{1}{T}\ell(\underline{\theta}) = \ell_\infty(\underline{\theta}_0) - \frac{1}{2}(\underline{\theta} - \underline{\theta}_0)' H(\underline{\theta}^*)(\underline{\theta} - \underline{\theta}_0)$$

since $-\lim E\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'}\Big|_{\underline{\theta}_0} = -\text{plim}\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'} = H(\underline{\theta})$.

$$\ell_\infty(\underline{\theta}) < \ell_\infty(\underline{\theta}_0) \quad for\ \underline{\theta} \neq \underline{\theta}_0$$

Choose a compact set $C \subseteq N(\underline{\theta}_0)$. *Then for any* $\underline{\theta}^* \in C$,

$$\lim_{T \to \infty} P[\frac{1}{T}\ell(\underline{\theta}^*) > \frac{1}{T}\ell(\underline{\theta})] = 1$$

which implies that

$$\lim_{T \to \infty} P[\underline{\theta}^* \in \Theta_0] = 1$$

where

$$\Theta_0 = \{\underline{\theta}; \frac{\partial \ell}{\partial \underline{\theta}} = \underline{0}\}$$

Therefore, the local ML estimator $\hat{\underline{\theta}}$ *converges in probability to* $\underline{\theta}_0$.

Asymptotic Normality:

Using a Taylor series expansion of $\frac{\partial \ell}{\partial \underline{\theta}} |_{\hat{\underline{\theta}}}$ *around* $\underline{\theta}_0$

$$\frac{\partial \ell}{\partial \underline{\theta}} |_{\hat{\underline{\theta}}} = \frac{\partial \ell}{\partial \underline{\theta}} |_{\underline{\theta}_0} + \frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'} |_{\underline{\theta}^*} (\hat{\underline{\theta}} - \underline{\theta}_0)$$

where $\|\underline{\theta}^* - \underline{\theta}_0\| \leq \|\hat{\underline{\theta}} - \underline{\theta}_0\|$. *Since* $\hat{\underline{\theta}} \in \Theta_T$, $\frac{\partial \ell}{\partial \underline{\theta}} |_{\hat{\underline{\theta}}} = 0$. *Therefore,*

$$\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) \stackrel{A}{=} -[\frac{1}{T} \frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'} |_{\underline{\theta}^*}]^{-1} \frac{1}{\sqrt{T}} \frac{\partial \ell}{\partial \underline{\theta}} |_{\underline{\theta}_0}$$

where $\stackrel{A}{=}$ *denotes asymptotic equivalence. According to assumption* (4)

$$-\text{plim} \frac{1}{T} \frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'} |_{\underline{\theta}^*} = H(\underline{\theta}_0)$$

since $\text{plim}\, \underline{\theta}^* = \underline{\theta}_0$. *Therefore,*

$$\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) \rightarrow N(\underline{0}, H(\underline{\theta}_0)^{-1} J(\underline{\theta}_0) H(\underline{\theta}_0)^{-1})$$

2.4.2 Maximum Likelihood Estimator

The likelihood function is defined by

$$L(\underline{\theta}; y_t, \underline{x}_t) = \prod_{t=1}^{T} f(\underline{\theta}, y_t, \underline{x}_t)$$

where y_t is the observed value of the random variable Y_t and $\underline{x}_t$ is an explanatory regressor vector at point t. Then $L(\underline{y}, X, \underline{\theta})$ is a density function with true parameter $\underline{\theta}_0$ satisfying

$$\int_{y_1} \cdots \int_{y_T} L(\underline{y}, X, \underline{\theta}_0) dy_1 \cdots dy_T = 1$$

In general, the density function $L(\underline{y}, X, \underline{\theta})$ is said to be regular for any $\underline{\theta} \in \Theta$ if differentiation with respect to $\underline{\theta}$ and subsequent integration over $\underline{y}$ yield the same result as in the reverse order. Therefore, if the density $L(\underline{\theta}, \underline{y}, X)$ is regular, the following is true:

$$E[\frac{\partial L(\cdot)}{\partial \underline{\theta}} / L(\cdot)] = 0 \tag{2.16}$$

$$E[\frac{\partial^2 L(\cdot)}{\partial \underline{\theta} \partial \underline{\theta}'} / L(\cdot)] = 0 \tag{2.17}$$

Consider the log-likelihood function $\ell = \log L(\underline{\theta}, \underline{y}, X)$. The score vector is

$$\underline{q}(\underline{\theta}) = \frac{\partial \ell}{\partial \underline{\theta}} = \frac{\partial L}{\partial \underline{\theta}} / L(\cdot)$$

Using this notation, equation (2.16) can be written as

$$E\underline{q}(\underline{\theta}) = 0$$

Then the Hessian matrix of $\ell(\cdot)$ is

$$\frac{\partial \underline{q}}{\partial \underline{\theta}'} = [\frac{\partial^2 L}{\partial \underline{\theta} \partial \underline{\theta}'} / L(\cdot) - \frac{\partial L}{\partial \underline{\theta}} \frac{\partial L}{\partial \underline{\theta}'} / L^2(\cdot)] \tag{2.18}$$

The Fisherian information matrix is defined as:

$$\Im(\underline{\theta}_0) = -E[\frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'} \mid_{\underline{\theta}_0}]$$

From Equations (2.17) and (2.18), it can be shown that

$$\Im(\underline{\theta}_0) = E\underline{q}(\underline{\theta}_0)\underline{q}(\underline{\theta}_0)' = V(\underline{q}(\underline{\theta}_0))$$

ML estimators have the desirable asymptotic properties — consistency and asymptotic normality. In addition,

$$\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) \sim N(\underline{0}, \lim_{T\to\infty}[-E\frac{1}{T}\frac{\partial^2 \ell}{\partial\underline{\theta}\partial\underline{\theta}'}|_{\underline{\theta}_0}^{-1})]$$

which means that $V(\hat{\underline{\theta}})$ is approximated by $[-E\frac{\partial^2 \ell}{\partial\underline{\theta}\partial\underline{\theta}'}|_{\underline{\theta}_0}]^{-1}$ which is the asymptotic Cramer-Rao lower bound. Since the asymptotic covariance matrix is a function of unknown parameters, the ML estimator of the asymptotic covariance matrix is given by:

$$\hat{V}_1 = [-E\frac{\partial^2 \ell}{\partial\underline{\theta}\partial\underline{\theta}'}|_{\hat{\underline{\theta}}}]^{-1} = [E\frac{\partial \ell}{\partial\underline{\theta}}\frac{\partial \ell}{\partial\underline{\theta}'}|_{\hat{\underline{\theta}}}]^{-1} \tag{2.19}$$

But the expected value of the Hessian is sometimes difficult to derive. Therefore, approximations for the inverted information matrix can be used as an alternative ML estimator if they are consistent:

$$\hat{V}_2 = [-\frac{\partial^2 \ell}{\partial\underline{\theta}\partial\underline{\theta}'}|_{\hat{\underline{\theta}}}]^{-1} \tag{2.20}$$

$$\hat{V}_3 = [\sum_{t=1}^{T}\frac{\partial \ell_t}{\partial\underline{\theta}}\frac{\partial \ell_t}{\partial\underline{\theta}'}|_{\hat{\underline{\theta}}}]^{-1} \tag{2.21}$$

In the context of small samples, however, these approximately equivalent covariance matrices may yield quite different results [Griffiths, Hill and Pope (1987); Calzorari and Panattoni (1988)]. Accordingly, the choice among these asymptotically equivalent covariances needs to be made with care rather than with the consideration of computational convenience, especially in finite samples.

2.4.3 Concentrated Maximum Likelihood Estimation

The method of concentrating the likelihood function makes it possible to exclude nuisance parameters without affecting the value of the estimated parameters. Consider the partitioned vector of parameters as $\underline{\theta} = [\underline{\beta}', \underline{\gamma}']'$ where $\underline{\gamma}$ is the nuisance parameter vector in which we are not interested. The first order condition for the maximization gives

$$\frac{\partial \ell}{\partial \underline{\beta}} |_{\underline{\hat{\beta}}} = \underline{0} \tag{2.22}$$

$$\frac{\partial \ell}{\partial \underline{\gamma}} |_{\underline{\hat{\gamma}}} = \underline{0} \tag{2.23}$$

The partitioned asymptotic information matrix is given by

$$H = -\lim E\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\theta}\partial \underline{\theta}'} = -\lim \begin{bmatrix} E\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\beta}\partial \underline{\beta}'} & E\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\beta}\partial \underline{\gamma}'} \\ E\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\gamma}\partial \underline{\beta}'} & E\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\gamma}\partial \underline{\gamma}'} \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

The required condition for the existence of concentrated likelihood estimators is that equation (2.23) be solved to produce an explicit representation of $\underline{\gamma}$ in terms of $\underline{\beta}$. Let $\underline{\gamma} = \underline{\hat{\gamma}}(\beta)$. Then we define the concentrated likelihood function:

$$\ell^c(\underline{\beta}) = \ell(\underline{\beta}, \underline{\hat{\gamma}}(\beta)) \tag{2.24}$$

Differentiating equation (2.23) with respect to $\underline{\beta}$, we obtain

$$\frac{\partial^2 \ell}{\partial \underline{\beta}\partial \underline{\gamma}'} |_{\underline{\theta}_0} + A(\underline{\beta})'[\frac{\partial^2 \ell}{\partial \underline{\gamma}\partial \underline{\gamma}'}_{\underline{\theta}_0}] = \underline{0}$$

where $A(\underline{\beta}) = \frac{\partial \underline{\hat{\gamma}}}{\partial \underline{\beta}'} |_{\underline{\beta}_0}$. Therefore,

$$A(\underline{\beta}) = -[\frac{\partial^2 \ell}{\partial \underline{\gamma}\partial \underline{\gamma}'}]^{-1}_{\underline{\theta}_0} \frac{\partial^2 \ell}{\partial \underline{\gamma}\partial \underline{\beta}'} |_{\underline{\theta}_0}$$

or

$$A(\hat{\underline{\beta}}) = -\left[\frac{\partial^2 \ell}{\partial \underline{\gamma} \partial \underline{\gamma}'}\right]^{-1}_{\hat{\underline{\theta}}} \frac{\partial^2 \ell}{\partial \underline{\gamma} \partial \underline{\beta}'} \Big|_{\hat{\underline{\theta}}} \tag{2.25}$$

The differentiation of both sides of (2.24) with respect to $\underline{\beta}$ leads to

$$\frac{\partial \ell^c}{\partial \underline{\beta}} \Big|_{\underline{\beta}_0} = \frac{\partial \ell}{\partial \underline{\beta}} \Big|_{\beta_0, \hat{\underline{\gamma}}} + A(\underline{\beta})' \frac{\partial \ell}{\partial \underline{\gamma}} \Big|_{\beta_0, \hat{\underline{\gamma}}} \tag{2.26}$$

Since $\frac{\partial \ell}{\partial \underline{\gamma}} \Big|_{\hat{\underline{\gamma}}} = \underline{0}$ from (2.23), equation (2.26) becomes

$$\frac{\partial \ell^c}{\partial \underline{\beta}} \Big|_{\underline{\beta}_0} = \frac{\partial \ell}{\partial \underline{\beta}} \Big|_{\beta_0, \hat{\underline{\gamma}}} \tag{2.27}$$

Differentiating equation (2.27) again with respect to $\underline{\beta}$, we obtain

$$\frac{\partial^2 \ell^c}{\partial \underline{\beta} \partial \underline{\beta}'} \Big|_{\underline{\beta}_0} = \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\beta}'} \Big|_{\beta_0, \hat{\underline{\gamma}}} + A(\underline{\beta}_0)' \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\gamma}'} \Big|_{\beta_0, \hat{\underline{\gamma}}} \tag{2.28}$$

Substituting (2.25) into (2.28)

$$\frac{\partial^2 \ell^c}{\partial \underline{\beta} \partial \underline{\beta}'} \Big|_{\underline{\beta}_0} = \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\beta}'} \Big|_{\beta_0, \hat{\underline{\gamma}}} - \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\gamma}'} \Big|_{\hat{\underline{\theta}}} \left[\frac{\partial^2 \ell}{\partial \underline{\gamma} \partial \underline{\gamma}'}\right]^{-1} \frac{\partial^2 \ell}{\partial \underline{\gamma}' \partial \underline{\beta}} \Big|_{\underline{\beta}_0, \hat{\underline{\gamma}}} \tag{2.29}$$

Therefore,

$$-\operatorname{plim} \frac{1}{T} \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\beta}'} \Big|_{\underline{\beta}_0} = H_{22} - H_{21} H_{22} H_{12} \tag{2.30}$$

By a Taylor series expansion around $\underline{\gamma}_0$,

$$\frac{\partial \ell^c}{\partial \underline{\beta}} \Big|_{\underline{\beta}_0} = \frac{\partial \ell}{\partial \underline{\beta}} \Big|_{\underline{\theta}_0} + \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\gamma}'} \Big|_{\underline{\theta}^*} (\hat{\underline{\gamma}} - \underline{\gamma}_0)$$

where $\|\underline{\theta}^* - \underline{\theta}_0\| \leq \|\overline{\underline{\theta}} - \underline{\theta}_0\|$ with $\overline{\underline{\theta}} = (\underline{\beta}^0, \hat{\underline{\gamma}})$. Let $\stackrel{A}{=}$ denote asymptotic equivalence. Using the relationship which comes from the asymptotic normality of $\hat{\underline{\theta}}$,

$$\sqrt{T}(\hat{\underline{\gamma}} - \underline{\gamma}_0) \stackrel{A}{=} H_{22}^{-1} \frac{1}{\sqrt{T}} \frac{\partial \ell}{\partial \underline{\gamma}} \Big|_{\underline{\theta}_0}$$

we can infer that

$$\frac{1}{\sqrt{T}}\frac{\partial \ell^c}{\partial \underline{\beta}} \mid_{\underline{\beta}_0} \stackrel{A}{=} [I, -\frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\gamma}'} \mid_{\underline{\theta}_0} H_{22}^{-1}] \frac{1}{\sqrt{T}}\frac{\partial \ell}{\partial \underline{\theta}} \mid_{\underline{\theta}_0}$$

Consequently,

$$\frac{1}{\sqrt{T}}\frac{\partial \ell^c}{\partial \underline{\beta}} \mid_{\underline{\beta}_0} \rightarrow N(\underline{0}, H_{11} - H_{21}H_{22}^{-1}H_{12}) \tag{2.31}$$

In the context of the concentrated likelihood function, the asymptotic covariance matrix is easily derived from (2.30) and (2.31):

$$V(\hat{\beta}) = [H_{11} - H_{21}H_{22}^{-1}H_{12}]^{-1}$$

This is exactly the submatrix of H^{-1} corresponding to the estimator $\underline{\hat{\beta}}$. Therefore, the estimator from the concentrated likelihood method is the same as that from the full ML estimation in the asymptotic sense.

2.4.4 General Equality Restrictions and Asymptotically Equivalent Test Statistics

Constrained Maximum Likelihood Estimation

The general equality restriction is written in the form:

$$g_i(\underline{\theta}) = 0, \quad i = 1, \ldots J$$

For ML estimation, the Lagrange multiplier method is used:

$$\ell(\underline{\theta}; \underline{y}, X) - \underline{g}'(\underline{\theta})\underline{\mu} \tag{2.32}$$

where $\underline{\mu} \in \Re^J$ is a Lagrange multiplier vector and $\underline{g} = (g_1, \ldots, g_J)'$ is a $J \times 1$ vector of constraint functions. For our analysis, the following assumptions about probability density functions and constraint functions are made:

1. There exist $\frac{\partial \ell}{\partial \theta_i}$, $\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}$ and $\frac{\partial^3 \ell}{\partial \theta_i \partial \theta_j \partial \theta_k}$ $(i,j,k = 1, \ldots, K)$ for every $\theta \in \mathrm{N}_\alpha = \{\underline{\theta} : \|\underline{\theta} - \underline{\theta}_0\| \leq \alpha\}$ and the first and second derivatives are continuous functions of $\underline{\theta}$.

2. For every $\underline{\theta} \in \mathrm{N}_\alpha$,
$$\begin{aligned} \|\tfrac{\partial \ell}{\partial \theta_i}\| &< G_1, \quad \forall i \\ \|\tfrac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\| &< G_2, \quad \forall j \\ \|\tfrac{\partial^3 \ell}{\partial \theta_i \partial \theta_j \partial \theta_k}\| &< G_3, \quad \forall k \end{aligned}$$
where G_1, G_2 and G_3 are integrable in $\Re$.

3. $E(\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j})$ is finite for $i,j = 1, \ldots, K$ and the matrix $H = -\lim E(\frac{1}{T} \frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'})$ is positive definite.

4. For every $\theta \in \mathrm{N}_\alpha$, $\frac{\partial g_m}{\partial \theta_i}$, $i = 1, \ldots, K$ and $m = 1, \ldots, J$ exist and these are continuous functions of $\underline{\theta}$.

5. For every $\theta \in \mathrm{N}_\alpha$, $\frac{\partial^2 g_m}{\partial \theta_i \partial \theta_j}, i,j = 1, \ldots, K, m = 1, \ldots, J$ exists and
$$\|\frac{\partial^2 g_m}{\partial \theta_i \partial \theta_j}\| < M,$$
where M is a given constant.

6. The $K \times J$ matrix $G = \frac{\partial \underline{g}'}{\partial \underline{\theta}}$ is of rank J.

Aitchison and Silvey (1958) proved the existence of a constrained ML estimator satisfying the above conditions. The constrained ML estimator $\tilde{\underline{\theta}}$ is obtained by solving the following equations:

$$\underline{q}(\tilde{\underline{\theta}}) - G\tilde{\underline{\mu}} = \underline{0} \tag{2.33}$$

$$\underline{g}(\tilde{\underline{\theta}}) = \underline{0} \tag{2.34}$$

where $\underline{q} = \frac{\partial \ell}{\partial \underline{\theta}}$. By a Taylor series expansion of (2.33) and (2.34) about $\underline{\theta}_0$, we can obtain the relationship

$$\begin{aligned} \frac{1}{\sqrt{T}}\underline{q}(\underline{\theta}_0) + \frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\theta}\partial \underline{\theta}'}\Big|_{\underline{\theta}_0} \sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) - \frac{1}{\sqrt{T}}G(\underline{\theta}_0)\tilde{\underline{\mu}} &= \underline{0} \\ \sqrt{T}G(\underline{\theta}_0)'(\tilde{\underline{\theta}} - \underline{\theta}_0) &= \underline{0} \end{aligned}$$

Therefore,

$$\begin{bmatrix} H(\underline{\theta}_0) & G(\underline{\theta}_0) \\ G(\underline{\theta}_0)' & 0_{J\times J} \end{bmatrix} \begin{bmatrix} \sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) \\ \frac{1}{\sqrt{T}}\tilde{\mu} \end{bmatrix} \underline{\underline{A}} \begin{bmatrix} \frac{1}{\sqrt{T}}\underline{q}(\underline{\theta}_0) \\ \underline{0} \end{bmatrix}$$

where $H(\underline{\theta}_0) = -\lim E\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\theta}\partial \underline{\theta}'}\Big|_{\underline{\theta}_0}$. Since the matrix

$$\begin{bmatrix} H(\underline{\theta}_0) & G(\underline{\theta}_0) \\ G(\underline{\theta}_0)' & 0 \end{bmatrix}$$

is nonsingular [Aitchison and Silvey (1958), pp. 822–3],

$$\begin{bmatrix} \sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) \\ \frac{1}{\sqrt{T}}\tilde{\mu} \end{bmatrix} \underline{\underline{A}} \begin{bmatrix} H(\underline{\theta}_0) & G(\underline{\theta}_0) \\ G(\underline{\theta}_0)' & 0_{J\times J} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\sqrt{T}}\underline{q}(\underline{\theta}_0) \\ \underline{0} \end{bmatrix}$$

Using the fact that

$$\frac{1}{\sqrt{T}}\underline{q}(\underline{\theta}_0) \to N(\underline{0}, -\lim E\frac{1}{T}\frac{\partial \ell}{\partial \underline{\theta}}\frac{\partial \ell}{\partial \underline{\theta}'}\Big|_{\underline{\theta}_0})$$

it follows that

$$\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) \to N(\underline{0}, H^{-1}[I - G(G'H^{-1}G)^{-1}G'H^{-1}])$$

Wald, Likelihood Ratio and Lagrange Multiplier Test

We consider the null hypothesis

$$H_0 : \underline{g}(\underline{\theta}) = \underline{0}$$

where $\underline{g}(\underline{\theta})$ is a $J \times 1$ vector-valued differentiable function. The Wald test is defined by

$$\mathrm{W} = \underline{g}(\hat{\underline{\theta}})'[\frac{\partial \underline{g}}{\partial \underline{\theta}'} \mid_{\hat{\underline{\theta}}} \hat{V}(\hat{\underline{\theta}}) \frac{\partial \underline{g}'}{\partial \underline{\theta}} \mid_{\hat{\underline{\theta}}}]^{-1} \underline{g}(\hat{\underline{\theta}})$$

where $\hat{\underline{\theta}}$ is a full ML estimator and $\hat{V}(\hat{\underline{\theta}})$ is a consistent estimator for the covariance matrix of $\hat{\underline{\theta}}$. The Wald test is more convenient than the likelihood ratio test since we need not estimate the constrained model. When the null hypothesis H_0 is correct, $\underline{g}(\hat{\underline{\theta}})$ should be close to $\underline{g}(\underline{\theta}_0)$. Therefore, we can reject the restriction if $\underline{g}(\hat{\underline{\theta}})$ is far from the null vector. The validity of a χ^2-approximation to the Wald test depends on the consistency of the covariance matrix estimator.

Nelson and Savin (1988) analyzed the power functions of the W, LM and LR tests for three classes of nonlinear models (nonlinear regression models with normal disturbances, logit models, and Tobit models) in small samples. They showed that the finite sample power functions of the Wald tests in one parameter models are nonmonotonic due to the explosive asymptotic standard errors of the estimate.

Gregory and Veall (1985) showed in their Monte Carlo simulation that the transformation from one nonlinear restriction into another algebraically equivalent form yielded a different test statistic in small samples. According to their

investigation, the null hypothesis $H_0^A : \beta_1\beta_2 - 1 = 0$ was indicative of some statistical advantage over another null hypothesis $H_0^B : \beta_1 - \frac{1}{\beta_2} = 0$. Phillips and Park (1988) studied the small sample behaviour of the Wald statistics under several nonlinear restrictions. They employed the Edgeworth expansion technique to obtain the finite distribution of these Wald statistics. They found a substantial gap between the asymptotic and finite sample distributions under $H_0^B : \beta_1 - \frac{1}{\beta_2} = 0$, which confirmed the Monte Carlo results of Gregory and Veall. Phillips and Park concluded that the small sample distribution of the Wald statistic might depend heavily on a different representation of the algebraically equivalent nonlinear restrictions.

The Lagrange Multiplier test can be derived from constrained ML estimators. Aitchison and Silvey (1958) and Silvey (1959) proposed the Lagrange multiplier test which is equivalent to Rao's score statistic:

$$\begin{aligned} \text{LM} &= \frac{1}{T}\tilde{\underline{\mu}}'[G(\tilde{\underline{\theta}})'H(\tilde{\underline{\theta}})G(\tilde{\underline{\theta}})]\tilde{\underline{\mu}} \\ &= -\underline{g}(\tilde{\underline{\theta}})'[\frac{\partial^2 \ell}{\partial\underline{\theta}\partial\underline{\theta}'}]_{\tilde{\underline{\theta}}}^{-1}\underline{g}(\tilde{\underline{\theta}}) \end{aligned}$$

The basic idea of the Lagrange multiplier test is that $\underline{g}(\tilde{\underline{\theta}})$ should be close to $\underline{g}(\hat{\underline{\theta}})$ if the null hypothesis is true. Since the Lagrange multiplier principle only requires constrained ML estimators, it is convenient to apply if we can reduce the number of parameters explicitly. Godfrey (1978a, 1978b) employed this principle to test ARMA procedures in the errors of dynamic regression equations. Davidson and MacKinnon (1985) provided a new test which followed the Lagrange multiplier

principle in testing the functional form of the model; namely, linear and log-linear regression against Box-Cox alternatives. Their test statistic showed good performance in small samples.

The likelihood ratio test is based on the likelihood function:

$$\text{LR} = -2(\ell(\underline{\tilde{\theta}}) - \ell(\underline{\hat{\theta}}))$$

Under the null hypothesis, the unconstrained ML value should not be far from the constrained ML value and thus LR should be small.

In the linear regression model, the Wald test statistic under the linear restriction $R\underline{\beta} = \underline{r}$ is given by

$$W = T(R\underline{b} - \underline{r})'[R(X'X)^{-1}R']^{-1}(R\underline{b} - \underline{r})/s$$

where $s = (\underline{y} - X\underline{b})'(\underline{y} - X\underline{b})$. The likelihood and Lagrange multiplier statistic can be written as a function of the Wald statistic:

$$\begin{aligned} LM &= W/(1 + W/T) \\ LR &= T\ln(1 + W/T) \end{aligned}$$

Therefore, the three asymptotically equivalent test statistics have the inequality relation: $LM \leq LR \leq W$. When using an asymptotic test procedure, the critical value for an α-level test based on the Wald, Lagrange multiplier or likelihood ratio statistic is approximated by $\chi^2_{(J)}(\alpha)$, where J is the number of restrictions. Therefore, the inequality relationship may lead to conflicting inference in linear

models [Berndt and Savin (1977); Evans and Savin (1982)]. The significance level can be corrected by using an Edgeworth expansion. For example, Rothenberg (1984) obtained size-corrected critical values in the linear regression model with nonscalar covariance error disturbances and showed that the power of the three tests is the same to order T^{-1}. In the context of nonlinear models, the small sample distribution of the three statistics is not known, and requires further research.

2.5 Small Sample Theory

In econometrics, statistical inference heavily depends on asymptotic theory since finite sample properties of econometric estimators and test statistics are rarely known. But the asymptotic properties are sometimes misleading when sample sizes are relatively small and the asymptotic procedures do not reflect the finite sample behaviour.

In the early stages of theoretical work on small sample distributions, the exact density functions for the two stage least squares and OLS estimators in simultaneous equations were developed. Two seminal articles [Basmann (1961); Bergstrom (1962)] provided an incentive for the development of the exact sample distribution theory for econometric estimators and test statistics. Basmann (1974) provided a survey of the literature in the area of finite sample theory in its early stages. Phillips (1980, 1985) and Rothenberg (1983) gave good descriptions of recent developments in this field.

2.5.1 Gram-Charlier and Edgeworth Expansion

Consider a sum of T independent random variables

$$S_T = X_1 + \cdots + X_T$$

where X_t has mean μ_t and standard deviation σ_t. Then the standardized variable $\frac{S_T - \sum_{t=1}^{T} \mu_t}{\sqrt{\sum_{t=1}^{T} \sigma_t^2}}$ is approximately distributed as $N(0,1)$ as T goes to ∞, according to the central limit theorem.

Theorem 2.3 Central Limit Theorem (Lindeverg and Lévy)

Let $\{X_t\}$ be independent and identically distributed with $E(X_t) = \mu$ and $V(X_t) = \sigma^2$. Then $Z_T \overset{a}{\sim} N(0,1)$ where

$$Z_T = \frac{S_T - T\mu}{\sqrt{T}\sigma}$$

Theorem 2.4 Central Limit Theorem (Liapounov)

Let $\{X_t\}$ be independent random variables with $E(X_t) = \mu_t$ and $V(X_t) = \sigma_t^2$. Suppose that the third absolute moment of $\{X_t\}$ about its mean μ_t is finite for every t. Let $m_{3t} = E|X_t - \mu_t|^3$. If the condition

$$\lim_{T \to \infty} (\sum_{t=1}^{T} m_{3t})^{1/3} (\sum_{t=1}^{T} \sigma_t^2)^{-1/2} = 0$$

is satisfied, then $Z_T \overset{a}{\sim} N(0,1)$ where

$$Z_T = \frac{S_T - \sum_{t=1}^{T} \mu_t}{\sqrt{\sum_{t=1}^{T} \sigma_t^2}}$$

The central limit theorem justifies the statement that the standardized sum of the independent variables has an approximate standard normal distribution $\Phi(\cdot)$. In addition, the density function of this standardized sum will be approximately equal to the standard normal probability density function $\phi(\cdot)$ if all the components of $\{X_t\}$ have continuous distributions and certain regularity conditions are satisfied [Cramer (1946)]:

$$F(x) = \Phi(x) + R(x) \tag{2.35}$$

$$f(x) = \phi(x) + r(x) \tag{2.36}$$

where $R(x)$ and $r(x)$ are small for large T and $R'(x) = r(x)$. From a general point of view, we can consider a random variable X which is not necessarily a sum of independent variables but seems to be approximately normal. Now we consider two types of expansions which use the functional form (2.36) and attempt to expand the remaining terms.

First, the type A Gram-Charlier series will be discussed [Cramer (1946); Phillips (1980)]. Let us consider a statistic S_T obtained from a sample of size T and its density $f_T(x)$. Then we may consider an expansion of the form:

$$f_T = c_0\phi(x) + \frac{c_1}{1!}\phi^{(1)}(x) + \cdots \tag{2.37}$$

where the c_i are constant coefficients and $\phi^{(i)}(x)$ is the ith derivatives of $\phi(x)$. The coefficients c_i in equation (2.37) can be determined by multiplying through by the Hermite polynomial $H_i(x) = (-1)^i \frac{\phi^{(i)}(x)}{\phi(x)}$, under the assumption that the series of

(2.37) can be integrated term by term. Multiplying both sides of (2.37) by $H_i(x)$,

$$\begin{aligned} H_i(x)f(x)_T &= c_0\phi(x)H_i(x) + \cdots + \frac{c_i}{i!}\phi^{(i)}(x)H_i(x) + \cdots \\ &= c_0\, H_0H_i(x)\phi(x) + \cdots + \frac{c_i}{i!}(-1)^i H_i(x)H_i(x)\phi(x) + \cdots \end{aligned}$$

Integrating term by term

$$\int_{-\infty}^{+\infty} H_i(x)f_T(x)dx =$$
$$c_0 \int_{-\infty}^{+\infty} H_0(x)H_i(x)\phi(x)\,dx + \cdots + (-1)^i\frac{c_i}{i!}\int_{-\infty}^{+\infty} H_i(x)H_i(x)\phi(x)\,dx + \cdots$$

Using the following orthogonality of Hermite polynomials

$$\int_{-\infty}^{+\infty} H_m(x)H_n(x) = \begin{cases} n! & \text{for } m = n \\ 0 & \text{for } m \neq n \end{cases}$$

we obtain

$$c_i = (-1)^i \int_{-\infty}^{+\infty} H_i(x)f_T(x)\,dx \tag{2.38}$$

Consequently, assuming that the standardized variable has finite moments of all orders,

$$f_T(x) = \phi(x) + \frac{c_3}{3!}\phi^{(3)}(x) + \frac{c_4}{4!}\phi^{(4)}(x) + \frac{c_5}{5!}\phi^{(5)}(x) + \cdots \tag{2.39}$$

where c_i is given by (2.38), i.e. $c_3 = -m_3/\sigma^3$, $c_4 = m_4/\sigma^4$ with the notation

$$m_i = E[x - E(x)]^i$$

But it should be noted that the series (2.39) is convergent if the integral

$$\int_{-\infty}^{+\infty} \exp(x^2/4)f_T(x)dx$$

is convergent and $f_T(x)$ is of bounded variation in $\Re$.

Definition 2.29 *Let f be a real valued function defined on the interval $[c,d]$ and $c = x_0 < x_1 < \cdots < x_K = d$ be an arbitrary subdivsion of $[c,d]$. Then f is said to be of* bounded variation *over $[c,d]$ if*

$$\sup_{K \in \mathcal{I}^+} \sum_{i=1}^{K} |f(x_i) - f(x_{i-1})| < \infty$$

for all possible values of K.

Unfortunately, only a small class of density functions can admit the expansions of (2.39) under these restrictions.

On the other hand, the Edgeworth expansion succeeds in providing a good approximation with a finite number of terms. Edgeworth series have a finite number of terms when we neglect the terms which have a specified order of magnitude, i.e. $O(T^{-3/2})$. If we take order of magnitude $T^{-(K+1)/2}$, then the asymptotic Edgeworth expansion has the form [Phillips (1980)]:

$$f_T(x) = \phi(x)[1 + \sum_{j=1}^{K} P_j(x) T^{-j/2}] + R_T(x)$$

where $P_j(x)$ is a polynomial of degree $3j$ and $R_T(x) = O(T^{-(K+1)/2})$. Suppose the independent and identically distributed random variables $X_1, \ldots, X_T$ have the continuous density function f and

$$E(X_t) = 0, \; V(X_t) = 1 \quad \text{for all } t.$$

In addition, we assume that the $\{X_t\}$ have finite moments of order $K+2$. Define the cumulants (semi-invariants) to be the coefficients of the following cumulant function:

$$\log \psi(t) = \sum_{j=1}^{K} \frac{\kappa_j}{j!}(it)^j + o(T^K),$$

where $\psi(t)$ is the characteristic function of X_t and κ_j is the jth cumulant of f [Cramer (1946), pp 185–6]. For example, the expansions of κ_j are

$$\begin{aligned}
\kappa_1 &= E(X_t) = 0 \\
\kappa_2 &= V(X_t) = 1 \\
\kappa_3 &= E(X_t - E(X_t))^3 \\
\kappa_4 &= E(X_t - E(X_t))^4 - 3
\end{aligned}$$

The jth cumulant of the standardized statistic $Z_T = \frac{\sum_{t=1}^{T} X_t}{\sqrt{T}}$ is

$$\gamma_j = \kappa_j T^{(1-j/2)}, \quad \text{for } j = 1, \ldots, K+2.$$

The Gram-Charlier series is

$$f_T = \phi(x) - \frac{1}{3!}\frac{\gamma_3}{T^{1/2}}\phi^{(3)}(x) + \frac{1}{4!}\frac{\gamma_4}{T}\phi^{(4)}(x) - \frac{1}{5!}\frac{\gamma_5}{T^{3/2}}\phi^{(5)}(x) + \frac{1}{6!}\left(\frac{\gamma_6}{T^2} + \frac{10\gamma_3^2}{T}\right)\phi^{(6)}(x) + \cdots$$

where, e.g., $\kappa_1 = c_1$ and $\kappa_2 = c_2 - c_1^2$. By a simple manipulation, the Edgeworth series has the form:

$$\begin{aligned}
f_T &= \phi(x) - \frac{1}{3!}\frac{\gamma_3}{T^{1/2}}\phi^{(3)}(x) \\
&\quad + \frac{1}{T}\left[\frac{1}{4!}\gamma_4\phi^{(4)}(x) + \frac{1}{72}\gamma_3^2\phi^{(6)}(x)\right] + O(T^{-3/2}) \\
&= \phi(x)\{1 + \frac{\gamma_3}{6}(x^3 - 3x)T^{-1/2} \\
&\quad + \left[\frac{\gamma_4}{24}(x^4 - 6x^2 + 3) + \frac{\gamma_3^2}{72}(x^6 - 15x^4 + 45x^2 - 15)\right]T^{-1} \\
&\quad + O(T^{-3/2}) \qquad (2.40)
\end{aligned}$$

The expansion of the corresponding distribution function of Z_T can be obtained by integrating the series in (2.40) term by term.

The Edgeworth expansion tends to give poor approximations in the tails of the distribution. Phillips (1977) investigated the usefulness of the Edgeworth expansion in the context of a first order stochastic difference equation and showed the severe distortion of the tail area probabilities. On the other hand, the Edgeworth expansion can provide information about the regions where asymptotic theory gives a good approximation and those where it works poorly.

2.5.2 Saddlepoint Approximation

Daniels (1954, 1980) explored the saddlepoint approximation, an alternative to the Edgeworth expansion which had unsatisfactory performance in the tail area. The saddlepoint approximation is always positive and has the same order of maginitude of error as the first two terms of the Edgeworth expansion. Holly and Phillips (1979) used a saddlepoint approximation to the distribution of the k-class estimator.

We consider the structural equation:

$$\underline{y}_1 = \beta \underline{y}_2 + Z_1\gamma_1 + \underline{u} \tag{2.41}$$

where $\underline{y}_1$ and $\underline{y}_2$ are $T\times 1$ vectors of endogenous variables and Z_1 is a $T\times K_1$ matrix of exogenous variables. The reduced form of equation (2.41) is

$$[\underline{y}_1 \vdots \underline{y}_2] = Z_1[\pi_{11} \vdots \pi_{12}] + Z_2[\pi_{21} \vdots \pi_{22}] + [\underline{v}_1 \vdots \underline{v}_2]$$

where $Z = [Z_1 \vdots Z_2]$ is a $T \times (K_1 + K_2)$ matrix of exogenous variables and $V = [\underline{v}_1 \vdots \underline{v}_2]$ a matrix of reduced form disturbances in which each row is independent and identically distributed as $N(\underline{0}, \Omega)$ and

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix}$$

The k-class estimator for β is given as

$$\hat{\beta}_k = \frac{\underline{y}_2' A_k \underline{y}_1}{\underline{y}_2' A_k \underline{y}_2}$$

where $A_k = (1-k)[I - Z_1(Z_1'Z_1)^{-1}Z_1' - Z_2(Z_2'Z_2)^{-1}Z_2'] + Z_2(Z_2'Z_2)^{-1}Z_2'$. By transforming the covariance matrix Ω to the identity matrix, the transformed k-class estimator can be written as

$$\hat{\beta}_k^* = \sqrt{\frac{\omega_{22}}{\omega_{11} - \omega_{12}^2/\omega_{22}}}(\hat{\beta}_k - \frac{\omega_{12}}{\omega_{22}})$$

Setting $k = 1$, the 2SLS estimator of the transformed model is

$$\hat{\beta}_1^* = \sqrt{\frac{\omega_{22}}{\omega_{11} - \omega_{12}^2/\omega_{22}}}(\frac{\underline{y}_2' Z_2(Z_2'Z_2)^{-1}Z_2'\underline{y}_1}{\underline{y}_2' Z_2(Z_2'Z_2)^{-1}Z_2'\underline{y}_2} - \frac{\omega_{12}}{\omega_{22}})$$

Holly and Phillips (1979) obtained the saddlepoint approximation of the probability density function of $\hat{\beta}_1^*$:

$$\begin{aligned} h(x) &= \frac{1}{\sqrt{2\pi\mu}} K_2 \beta^{K_2} (\beta^2 - 2\beta x - 1)^{-(K_2+1)/2} e^{-\mu^2/2}, \text{ when } x < \tfrac{\beta^2-1}{2\beta} \text{ and } \beta > 0 \\ &= \frac{1}{\sqrt{2\pi\mu}} [K_2 + \mu^2 \frac{1+2\beta-\beta^2}{1+x^2}](\beta x + 1)^{K_2} (1 + 2\beta x - \beta^2)^{-(K_2+1)/2} \cdot \\ &\quad (1+x^2)^{-K_2/2} \exp[-\frac{\mu^2}{2}\frac{(x-\beta)^2}{1+x^2}], \text{ when } x > \tfrac{\beta^2-1}{2\beta} \text{ and } \beta > 0 \end{aligned}$$

where $\mu^2 = \frac{\pi_{22}'(Z_2'Z_2)\pi_{22}}{\omega_{22}}$. The distribution function which was obtained by numerical calculation indicated a uniformly better performance in the tail area than the Edgeworth expansion up to $O(\mu^{-1})$ and $O(\mu^{-3})$. In addition, the saddlepoint approximation showed less sensitivity to parameter variations than the Edgeworth expansion. But the saddlepoint technique is applicable only in the special case where the characteristic function is available or there is a set of sufficient statistics for the parameters.

2.5.3 Rational Function Approximation

As an alternative to deriving the finite distributions of estimators and test statistics, a method of approximating the small sample distributions was developed by Phillips (1982, 1983) using modified multiple-point Padé approximants to the distribution. This technique has advantages over the usual approximation methods (Edgeworth and saddle points) in that accurate approximations can be obtained in very small samples and *a priori* or additional information on the distribution can be incorporated. Phillips proved the existence, uniqueness and convergence of best approximating rational functions to continuous density functions. Phillips (1983) defined the extended family of rational approximating functions (ERA) as follows:

Definition 2.30 *Define the class of extended rational approximating functions of maximal degree m and n as the class of functions R_{mn} of the form:*

$$R_{mn}(x; s, \underline{\gamma}) = s(x)\frac{P_m}{Q_n} = s(x)\frac{\sum_{i=1}^{m} a_i x^i}{\sum_{i=1}^{n} b_i x^i},\ x \in \Re \text{ and } m \leq n$$

where $s(x)$ is a real continuous function, $P_m(x)$ and $Q_n(x)$ are reduced to their lowest degree cancelling identical factors, m and n are even integers and

$$\underline{\gamma} = (a_1, \ldots, a_m, b_1, \ldots, b_n)' \in \{\underline{\gamma} : \sum_{i=1}^{n} b_i^2 = 1, Q_n(x) > 1\ \forall x \in \Re\}$$

The additional information can be incorporated into the approximating function $R_{mn}(x)$ via $s(x)$ — for example, a probability density estimated from Monte Carlo simulation, a crude asymptotic approximation to the true density or the fact that the density is greater than zero. There is a possibility that this ERA family includes approximating functions which are singular and which have negative probability over some part of the parameter space, but these can be eliminated by restricting $P_m > 0$.

Consider the case of an approximation with $m = n = 4$. Let

$$[n/n](x) = \frac{P_n(x)}{Q_n(x)} = \frac{\sum_{i=1}^{n} a_i x^i}{\sum_{i=1}^{n} b_i x^i}$$

Then the Padé approximation function is given with points of local Taylor expansions at $x = 0$ and $x^{-1} = 0$:

$$[4/4](x) = \frac{P_4}{Q_4} = \frac{\alpha_4(x - \gamma_1)(x - \gamma_2)(x - \gamma)(x - \bar{\gamma})}{\beta_4(x - \delta_1)(x - \delta_2)(x - \delta)(x - \bar{\delta})} \tag{2.42}$$

where (γ_1, γ_2) and (δ_1, δ_2) denote the real zeros of the numerator and denominator, respectively, and $(\gamma, \overline{\gamma})$ and $(\delta, \overline{\delta})$ are complex conjugate pairs. We can modify (2.42) so that $[4/4](x) > 0, \forall x \in \Re$:

$$[4/4](x;\underline{\theta}) = \frac{\alpha_4[ax^2 + cx + c](x-\gamma)(x-\overline{\gamma})}{\beta_4[ex^2 + fx + g](x-\delta)(x-\overline{\delta})} \tag{2.43}$$

where $\underline{\theta} = (a, b, c, d, e, f)'$. To keep the equivalent local behaviour of (2.42) and (2.43), we need to place restrictions on the parameters:

$$\begin{aligned} a/d &= 1 \\ c/f &= \gamma_1\gamma_2/\delta_1\delta_2 \end{aligned}$$

With these restrictions, it is ensured that

$$\begin{aligned} [4/4](x;\underline{\theta}) &\rightarrow \alpha_4/\beta_4 && \text{as } x \rightarrow \pm\infty \text{ and} \\ [4/4](x;\underline{\theta}) &\rightarrow \alpha_4\gamma_1\gamma_2|\gamma|^2/\beta_4\delta_1\delta_2|\delta|^2 && \text{as } x \rightarrow 0 \end{aligned}$$

Using ERA, there is a tradeoff between reliable global behaviour and good local performance. Therefore, we can improve the global performance by trading off error magnitudes in different regions of the parameter space.

CHAPTER 3
GENERAL TRANSFORMATION-OF-VARIABLES IN REGRESSION

3.1 Introduction

Zarembka (1974) introduced the transformation of variables technique to econometrics. He explained and extended the model of Box and Cox (1964),

$$y_t^{(\lambda)} = \underline{x}_t'\underline{\beta} + \epsilon_t$$

where

$$y_t^{(\lambda)} = \begin{cases} (y_t^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \ln(y_t) & \text{if } \lambda = 0 \end{cases}$$

and

$$\epsilon_t \sim \text{iid} \quad N(0, \sigma^2)$$

Since economic theory usually fails to provide economists with useful guidance in specifying the functional form of an economic relationship, Zarembka suggested the transformation-of-variables model for choosing the appropriate functional form within the framework of statistical inference. The beauty of the general Box-Cox model is summarized in three ways. First, assuming that the dependent variable differs from the theoretical model by an additive random disturbance is often a misspecification. The restriction to a specific functional form loosely based on economic theory may lead to incorrect conclusions by ignoring sample information. Therefore, the functional form often needs to be estimated, even though a

functional relationship exists *a priori.* Second, the misspecification of the dependent variable can be avoided when an appropriate dependent variable is chosen from among those that are functionally related. Finally, the properties of the Box-Cox transformation (additive effects, constant error variance, and normally distributed error disturbance) provide the possibility of using the classical linear regression model when the transformation parameters are used. The risk of selecting one functional form can be reduced by testing the alternative functional forms that are estimated by use of OLS under the assumed power parameter values.

There exist many applications of the Box-Cox transformation in an econometric context. The typical application is found in the estimation of demand equations. Zarembka (1968) estimated the functional form of the money demand equation with the same power transformation applied to both the dependent and independent variables. White (1972) also estimated the money demand equation with the same functional parameters in the dependent and independent variables, among which the interest rate variable had a shifted location. Spitzer (1976) extended White's study by applying a general transformation-of-variables model to the money demand equation. The study of the U.K. money demand equation by Mills (1978) and Boylan and O'Muircheartaigh (1981) added to this literature. Also, the import demand equation was studied by Khan and Ross (1977), Hwang (1981), Boylan et al (1982), and Blaylock and Smallwood (1985) using the Box-Cox transformation technique. The Box-Cox transformation was also used in

estimating hedonic price models [Blackley et al (1984); Megbolugbe (1986)]. More examples can be listed in a wide range of areas [Granger and Newbold (1976); Hopwood et al (1984); Smyth and Dua (1986); Guerrero (1987); Montmarquette and Blais (1987)].

The asymptotic properties of the ML estimators of the general Box-Cox model include consistency, asymptotic unbiasedness, normality, and efficiency if the maximum likelihood regularity conditions are satisfied. But the density function of the dependent variable is not regular because the dependent variable is truncated unless the transformation parameter of the dependent variable is equal to zero [Hinkley (1975)]. Therefore, the usual asymptotic properties of the ML estimators may not be applied to the Box-Cox ML estimators in a strict sense. Furthermore, finite sample properties of the ML estimators of the Box-Cox transformation are not known. Spitzer (1978) investigated the small sample properties of the ML estimators when both the dependent and independent variables are transformed with the same power parameter. But Spitzer's Monte Carlo simulation has too small a number of replicates (50) for studying sampling behaviour and his model is too restrictive in employing the same power transformation for the dependent and independent variables.

This study includes an extensive Monte Carlo simulation designed to explore the unknown properties of the ML estimators of the general Box-Cox model and their finite sample variabilities. In Section 3.2, alternative estimation procedures

are outlined. It is a practical convention to test the significance of a coefficient using an asymptotic t ratio or squared t ratio (Wald statistic). Therefore, the Monte Carlo simulation will also be used to investigate the finite sample distribution of the t ratio. The exact distribution of the t statistic can be well approximated using the Edgeworth expansion, thus leading to improved confidence intervals. The Edgeworth expansion of the asymptotic t ratio requires information about the asymptotic variance and its first and second derivatives with respect to the corresponding parameter estimator. Unfortunately, the asymptotic variance and its derivatives for the nonlinear ML estimator in the Box-Cox model are difficult to obtain or may not exist, even if the error term is normally distributed. Therefore, the bootstrap inversion of an Edgeworth expansion is discussed in Section 3.3. In Section 3.4, the design of the Monte Carlo study is given and in Section 3.5 the results are presented. Section 3.6 contains a summary and concluding remarks.

3.2 Estimation of the Box-Cox Transformation

Box and Cox (1964) employed the ML method to estimate the parameters of their model. Spitzer (1982) discussed several estimation methods (full ML estimation, concentrated ML estimation, nonlinear least squares, and iterative OLS) in the context of the regression model when only the dependent variable is transformed. But nonlinear least squares applied to scaled variables is more heuristic than rigorous. In addition, iterative OLS is not practical for use in the general transformation-of-variables model and its covariance matrix estimator is not con-

sistent. The usual ML estimators are not consistent either and their asymptotic covariance matrix estimators (e.g., the inverse of the negative Hessian) are not correct because of the problem of truncation of the transformed dependent variable which is inherent in the Box-Cox model. Therefore, Amemiya and Powell (1981) proposed nonlinear two stage least squares, since this gives a consistent estimator if the error term has an expected value of zero, regardless of its distribution. We will consider full and concentrated ML, nonlinear two stage least squares, and iterative GLS estimation. Iterative GLS is considered because of its GLS interpretation and the fact that its only requirement is the first derivative of the log-likelihood function. In our Monte Carlo simulation, nonlinear two stage least squares is excluded since it may not give reasonable estimates due to the fact that all values of the dependent variable of our model are greater than one if $\lambda_1 > 0$; less than one if $\lambda_1 < 0$. This point will be discussed in detail in Section 3.2.2.

3.2.1 Maximum Likelihood Method

We can specify the generalized Box-Cox model in the form

$$y_t^{(\lambda_1)} = \beta_1 + \beta_2 x_{2t}^{(\lambda_2)} + \cdots + \beta_k x_{kt}^{(\lambda_k)} + \epsilon_t, \quad t = 1, \ldots, T \tag{3.1}$$

where

$$z_{it}^{(\lambda_i)} = \begin{cases} (z_{it}^{\lambda_i} - 1)/\lambda_i & \text{if } \lambda_i \neq 0 \\ \ln(z_{it}) & \text{if } \lambda_i = 0 \end{cases}$$

$$z_{it} = \begin{cases} y_t & \text{for } i = 1 \\ x_{it} & \text{for } i > 1 \end{cases}$$

$$\epsilon_t \sim \text{iid } N(0, \sigma^2)$$

Matrix algebra gives a more compact notation for Equation (3.1),

$$\underline{y}^{(\lambda_1)} = X^{(\lambda_\cdot)}\underline{\beta} + \underline{\epsilon} \tag{3.2}$$

where $\underline{y}^{(\lambda_1)}$ is a $T \times 1$ vector of the transformed dependent variable and $X^{(\lambda_\cdot)}$ is a $T \times k$ matrix of independent variables with the general Box-Cox transformation applied to each column vector except for the first, which is a constant vector. Let $\underline{\theta} = (\beta_1, \beta_2, \ldots, \beta_k, \lambda_1, \ldots, \lambda_k, \sigma^2)'$. The log-likelihood function for (3.2) is given by

$$\begin{aligned} \ell(\underline{\theta}; X, \underline{y}) = & -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 \\ & - \frac{1}{2\sigma^2}(\underline{y}^{(\lambda_1)} - X^{(\lambda_\cdot)}\underline{\beta})'(\underline{y}^{(\lambda_1)} - X^{(\lambda_\cdot)}\underline{\beta}) \\ & + (\lambda_1 - 1)\sum_{t=1}^{T} \ln y_t \end{aligned} \tag{3.3}$$

Consider the first order conditions for a maximum of the log-likelihood function (3.3):

$$\begin{aligned} \frac{\partial \ell}{\partial \underline{\beta}} &= \frac{1}{\sigma^2} X^{(\lambda_\cdot)'}\underline{\epsilon} = \underline{0} \\ \frac{\partial \ell}{\partial \lambda_1} &= -\frac{1}{\sigma^2}\underline{y}'_{\lambda_1}\underline{\epsilon} + \sum \ln y_t = 0 \end{aligned}$$

$$\begin{aligned}\frac{\partial \ell}{\partial \lambda_i} &= \frac{1}{\sigma^2}\beta_i \underline{x}'_{\lambda_i}\underline{\epsilon} = \underline{0}, \quad i > 1 \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{T}{2}\sigma^2 + \frac{1}{2\sigma^4}\underline{\epsilon}'\underline{\epsilon} = \underline{0}\end{aligned}$$

where

$$\underline{z}_{\lambda_i} = \frac{\partial \underline{z}_i^{(\lambda_i)}}{\partial \lambda_i} = \begin{cases} [(1+\lambda_i \underline{z}_i^{(\lambda_i)}) \# \ln(1+\lambda_i \underline{z}_i^{(\lambda_i)}) - \lambda_i \underline{z}_i^{(\lambda_i)}]/\lambda_i^2 & \text{if } \lambda_i \neq 0 \\ (\ln \underline{z}_i) \# (\ln \underline{z}_i)/2 & \text{if } \lambda_i = 0 \end{cases}$$

$$\underline{z}_i^{(\lambda_i)} = \begin{cases} \underline{y}^{(\lambda_1)} & \text{if } i = 1 \\ \underline{x}_i^{(\lambda_i)} & \text{if } i > 1 \end{cases}$$

and # denotes elementwise multiplication. Let

$$\underline{z}_{\lambda_i \lambda_i} = \frac{\partial \underline{z}_{\lambda_i}}{\partial \lambda_i}$$

Then the second derivatives of the log-likelihood function are

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\beta}'} &= -\frac{1}{\sigma^2} X^{(\lambda_\cdot)'} X^{(\lambda_\cdot)} \\ \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \lambda_1} &= -\frac{1}{\sigma^2} X^{(\lambda_\cdot)'} \underline{y}_{\lambda_1} \\ \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \lambda_i} &= -\frac{1}{\sigma^2}\left(\frac{\partial \beta_i}{\partial \underline{\beta}} \underline{x}'_{\lambda_i}\underline{\epsilon} - \beta_i X^{(\lambda_\cdot)'} \underline{x}_{\lambda_i}\right), \quad i = 2, \ldots, k \\ \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \sigma^2} &= -\frac{1}{\sigma^4} X^{(\lambda_\cdot)'} \underline{\epsilon} \\ \frac{\partial^2 \ell}{\partial \lambda_1^2} &= -\frac{1}{\sigma^2}(\underline{y}'_{\lambda_1}\underline{y}_{\lambda_1} + \underline{y}'_{\lambda_1 \lambda_1}\underline{\epsilon}) \\ \frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_i} &= \frac{1}{\sigma^2}\beta_i \underline{y}'_{\lambda_1}\underline{x}_{\lambda_i}, \quad i = 2, \ldots, k \\ \frac{\partial^2 \ell}{\partial \lambda_1 \partial \sigma^2} &= \frac{1}{\sigma^4}\underline{y}'_{\lambda_1}\underline{\epsilon} \\ \frac{\partial^2 \ell}{\partial \lambda_i^2} &= -\frac{1}{\sigma^2}(\beta_i^2 \underline{x}'_{\lambda_i}\underline{x}_{\lambda_i} - \beta_i \underline{x}'_{\lambda_i \lambda_i}\underline{\epsilon}), \quad j = 2, \ldots, k \\ \frac{\partial^2 \ell}{\partial \lambda_i \partial \lambda_j} &= -\frac{1}{\sigma^2}\beta_i \beta_j \underline{x}'_{\lambda_i}\underline{x}_{\lambda_j}, \quad i \neq j\end{aligned}$$

$$\frac{\partial^2 \ell}{\partial(\sigma^2)^2} = -\frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\underline{\epsilon}'\underline{\epsilon}$$

The information matrix of $\underline{\hat{\theta}}$ is written as

$$I(\underline{\hat{\theta}}) = -E[\frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'}]$$

$$= \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

where

$$I_{11} = -E \begin{bmatrix} \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\beta}'} & \frac{\partial^2 \ell}{\partial \underline{\beta} \partial \underline{\lambda}'} \\ \frac{\partial^2 \ell}{\partial \underline{\beta}' \partial \underline{\lambda}} & \frac{\partial^2 \ell}{\partial \underline{\lambda} \partial \underline{\lambda}'} \end{bmatrix}$$

$$I_{21} = -E(0, \ldots, 0, \frac{\partial^2 \ell}{\partial \lambda_1 \partial \sigma^2}, 0, \ldots, 0)$$

$$I_{22} = -E[\frac{\partial^2 \ell}{\partial(\sigma^2)^2}]$$

From the first order conditions,

$$\hat{\sigma}^2 = \frac{1}{T}(\underline{y}^{(\hat{\lambda}_1)} - X^{(\hat{\lambda}.)}\underline{\hat{\beta}})'(\underline{y}^{(\hat{\lambda}_1)} - X^{(\hat{\lambda}.)}\underline{\hat{\beta}}) \tag{3.4}$$

The full ML function can be concentrated as

$$\ell^c(\cdot) = constant - \frac{T}{2}\ln(\hat{\sigma}^2) + (\lambda_1 - 1)\sum \ln(y_t) \tag{3.5}$$

The parameters β_i and λ_i $(i = 1, \ldots, K)$ are estimated by maximizing the concentrated likelihood function (3.5), and subsequently $\hat{\sigma}^2$ is obtained using Equation (3.4).

Under the usual regularity conditions for the likelihood function, the ML estimators are consistent and asymptotically efficient.

$$\underline{\hat{\theta}} = \underline{\theta}_0 + o_P(T)$$

$$asy.V(\hat{\underline{\theta}}) \;=\; [I(\underline{\theta}_0)]^{-1}$$

where $I(\underline{\theta}_0)$ is the information matrix. But the asymptotic covariance matrix of the ML estimators of the Box-Cox parameters is difficult to derive because the expectation of the Hessian matrix cannot be evaluated analytically. Draper and Cox (1969) derived the variance of a power transformation estimator from the information matrix using a series expansion. They assumed that the transformed dependent variable $z = y^{(\lambda)}$ has a normal distribution. But from a theoretical point of view, the distribution of the original dependent variable is required in order to obtain a regular likelihood function [Hinkley (1975); Amemiya and Powell (1981)]. Hinkley (1975) employed a one parameter gamma density $f(y) = y^{\rho-1}e^{-y}/\Gamma(\rho)$ and Amemiya and Powell (1981) a two parameter gamma density $f(y) = \frac{1}{\Gamma(\rho)}\alpha^{-\rho}y^{\rho-1}e^{-(1/\alpha)y}$, while Poirier (1978) used the truncated normal density. But we will not make any assumption about the distribution of the original dependent variable since this study focuses on the finite properties of the Box-Cox model and on a Monte Carlo simulation instead of rigorous theoretical analysis. Therefore, the inverse of the negative Hessian evaluated at the ML estimator $\hat{\underline{\theta}}$ will be used as a maximum likelihood estimator of $V(\hat{\underline{\theta}})$ when we make the usual assumption that $-\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\theta}\partial \underline{\theta}'}\,|_{\hat{\underline{\theta}}}$ converges in probability to a finite nonsingular matrix $C_0 \;= -\lim E\frac{1}{T}\frac{\partial^2 \ell}{\partial \underline{\theta}\partial \underline{\theta}'}|_{\underline{\theta}_0}$.

Furthermore,

$$\sqrt{T}(\hat{\underline{\theta}} \;- \underline{\theta}_0) \rightarrow N(\underline{0}, C_0^{-1})$$

Therefore, it is straightforward to infer that

$$\frac{\hat{\theta}_i - \theta_{0i}}{\sqrt{V(\hat{\theta}_i)}} \rightarrow N(0,1)$$

By replacing $V(\hat{\theta}_i)$ by its consistent estimator, the asymptotic t ratio is obtained.

3.2.2 Nonlinear Two Stage Estimation (NL2SE)

Amemiya and Powell (1981) proposed an alternative estimation method for the Box-Cox transformation. To implement the NL2SE method, we modify model (3.2) as follows:

$$\underline{f}(\underline{y}, X;\ \underline{\beta}, \underline{\lambda}) = \underline{y}^{(\lambda_1)} - X^{(\lambda_\cdot)'}\underline{\beta} = \underline{\epsilon} \tag{3.6}$$

The minimand for NL2SE is

$$S(\underline{\beta}, \underline{\lambda};\ W) = \underline{f}'W(W'W)^{-1}W'\underline{f} \tag{3.7}$$

where W is a $T \times N$ matrix which is usually composed of quadratic functions of the columns of the X matrix as well as the X matrix itself. The NL2SE method should be applied with care since NL2SE is not well defined when the values of the dependent variable all exceed 1 or are less than 1 [Khazzoom (1989)]. Consider the original Box-Cox model:

$$\underline{y}^{(\lambda)} = X\underline{\beta} + \underline{\epsilon}$$

Given λ, $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}^{(\lambda)}$ is the estimator which minimizes

$$(\underline{y}^{(\lambda)} - X\underline{\beta})'W(W'W)^{-1}W'(\underline{y}^{(\lambda)} - X\underline{\beta})$$

Then the concentrated minimand is written as

$$\underline{y}^{(\lambda)'}(I - X(X'X)^{-1}X')W(W'W)^{-1}W'(I - X(X'X)^{-1}X')\underline{y}^{(\lambda)}$$

Suppose the W matrix contains the X matrix as $[X \vdots W_r]$ where W_r is the remainder of the W matrix after excluding the matrix X. Let $P_X = X(X'X)^{-1}X'$ and $P_W = W(W'W)^{-1}W'$. Then it is easy to show

$$P_W = P_X + (I - P_X)W_r[W_r'W_r - W_r'P_X W_r]^{-1}W_r'(I - P_X)$$

Using the fact that $(I - P_X)X = \underline{0}$ and $P_X X = X$, it follows that $P_W X = X$. Thus, the concentrated minimand can be written

$$S_P = \underline{y}^{(\lambda)'}(P_W - P_X)\underline{y}^{(\lambda)}$$

When all the elements of $\underline{y}$ are greater than 1, $S_P \to 0$ as $\lambda \to -\infty$. The opposite behaviour of S_P according to the change in λ is observed when all the values of $\underline{y}$ are less than 1. In other words, the moment condition that $E(\underline{y}^{(\lambda)} - X\underline{\beta})'W = 0$ is not sufficient to identify the true parameters.

3.2.3 Iterative Generalized Least Squares (IGLS)

Full information ML estimation in linear and nonlinear simultaneous equation models can be interpreted as iterative generalized least squares [Dagenais (1978)]. Similar procedures can be applied to single equation ML estimation. The main advantage of this method is that second derivatives are not necessary and nonlinear maximum likelihood estimators can be interpreted in the context of generalized

least squares. Consider the log-likelihood of the Equation (3.6):

$$\ell(\underline{\theta};\ \underline{y}, X) = constant + \frac{T}{2}\ln(\sigma^{-2}) - \frac{1}{2\sigma^2}(\underline{\epsilon}'\underline{\epsilon}) + \sum \ln|J_t| \tag{3.8}$$

where $J_t = \frac{\partial \epsilon_t}{\partial y_t} = \frac{\partial f_t}{\partial y_t} = y_t^{\lambda_1 - 1}$. The first order conditions for the maximization of the likelihood function are as follows:

$$\frac{\partial \ell}{\partial \theta_i} = \sum \frac{1}{|J_t|}\frac{\partial |J_t|}{\partial \theta_i} - \hat{\sigma}^{-2}\frac{\partial \hat{\underline{\epsilon}}'}{\partial \theta_i}\hat{\underline{\epsilon}} = \underline{0} \tag{3.9}$$

$$\frac{\partial \ell}{\partial \hat{\sigma}^2} = -\frac{T}{2}\hat{\sigma}^{-2} + (\hat{\underline{\epsilon}}'\hat{\underline{\epsilon}})\hat{\sigma}^{-4} = 0 \tag{3.10}$$

where θ_i corresponds to the elements of $\underline{\beta}$ and $\underline{\lambda}$. From Equation (3.10),

$$\frac{(\hat{\underline{\epsilon}}'\hat{\underline{\epsilon}})}{T}\hat{\sigma}^{-2} = 1$$

Thus

$$\begin{aligned}\frac{\partial \ell}{\partial \theta_i} &= \sum \frac{1}{|J_t|}\frac{\partial |J_t|}{\partial \theta_i}\frac{(\hat{\underline{\epsilon}}'\hat{\underline{\epsilon}})}{T}\hat{\sigma}^{-2} - \hat{\sigma}^{-2}\frac{\partial \hat{\underline{\epsilon}}'}{\partial \theta_i}\hat{\underline{\epsilon}} \\ &= \hat{\sigma}^{-2}[\sum \frac{1}{|J_t|}\frac{\partial |J_t|}{\partial \theta_i}\frac{\hat{\underline{\epsilon}}'}{T} - \frac{\partial \hat{\underline{\epsilon}}'}{\partial \theta_i}]\hat{\underline{\epsilon}}\end{aligned}$$

Then by stacking

$$[\sum_{t=1}^{T} \frac{1}{|J_t|}\frac{\partial |J_t|}{\partial \theta_i}\frac{\hat{\underline{\epsilon}}'}{T} - \frac{\partial \hat{\underline{\epsilon}}'}{\partial \theta_i}]$$

we obtain

$$V' = \begin{bmatrix} \sum \frac{1}{|J_t|}\frac{\partial |J_t|}{\partial \theta_1}\frac{\hat{\underline{\epsilon}}'}{T} - \frac{\partial \hat{\underline{\epsilon}}'}{\partial \theta_1} \\ \vdots \\ \sum \frac{1}{|J_t|}\frac{\partial |J_t|}{\partial \theta_p}\frac{\hat{\underline{\epsilon}}'}{T} - \frac{\partial \hat{\underline{\epsilon}}'}{\partial \theta_p} \end{bmatrix}$$

where $p(= 2k)$ is the number of parameters to be estimated. The first order condition (3.9) satisfying Equation (3.10) is written as

$$\frac{\partial \ell}{\partial \underline{\theta}} = \hat{\sigma}^{-2}V'\hat{\underline{\epsilon}} = \underline{0} \tag{3.11}$$

Equation (3.11) can be written as

$$\hat{\sigma}^{-2}V'(\hat{\underline{\epsilon}} + V\hat{\underline{\theta}}) - \hat{\sigma}^{-2}V'V\hat{\underline{\theta}} = \underline{0} \tag{3.12}$$

Then Equation (3.12) is solved iteratively by setting

$$\tilde{\underline{\theta}}_{i+1} = (V_i'V_i)^{-1}V_i'(\hat{\epsilon}_i + V_i\hat{\underline{\theta}}_i)$$

and

$$\begin{aligned}\hat{\underline{\theta}}_{i+1} &= \hat{\underline{\theta}}_i + \lambda_i(\tilde{\underline{\theta}}_{i+1} - \hat{\underline{\theta}}_i) \\ \delta_i &= \hat{\underline{\theta}}_{i+1} - \hat{\underline{\theta}}_i \\ &= \lambda_i[(V_i'V_i)^{-1}V_i'\hat{\underline{\epsilon}}_i]\end{aligned}$$

where $V_i'\hat{\underline{\epsilon}}_i$ is the maximization gradient of the log-likelihood function at the ith iteration and $(V_i'V_i)$ is positive definite. Using the derived matrix V, Equation (3.6) is given by

$$\begin{aligned}\underline{f}(\underline{y}, X; \underline{\beta}, \underline{\lambda}) + V\underline{\theta} &= V\underline{\theta} + \underline{\epsilon} \\ W^* &= V\underline{\theta} + \underline{\epsilon}\end{aligned} \tag{3.13}$$

Then the IGLS esimator can be obtained by applying the least squares technique to Equation (3.13) iteratively. Consequently, the IGLS estimator of θ is

$$\hat{\underline{\theta}}_{IGLS} = (V_L'V_L)^{-1}V_L'W_L^* \tag{3.14}$$

where V_L and W_L^* are values of V and W^* evaluated at the maximum of the likelihood function. The consistent estimator of the asymptotic covariance matrix

is

$$\hat{V}(\underline{\hat{\theta}}_{IGLS}) = \hat{\sigma}^2(V_L'V_L)^{-1} \tag{3.15}$$

Obviously, the parameter estimates $\underline{\hat{\theta}}_{IGLS}$ are expected to be equal to the estimates obtained by use of the Newton-Raphson or BHHH algorithms. But the covariance matrix estimator from Equation (3.15) may be different from other consistent covariance estimators such as the inverted negative Hessian matrix. Calzolary and Panattoni (1988) showed that the estimated standard errors from the IGLS covariance matrix were generally smaller than those from the inverse of the negative Hessian matrix.

3.3 Bootstrapping and Edgeworth Expansion

Currently, there is no theoretical foundation for statistical inference based on estimators from nonlinear models in the context of finite samples. Bootstrapping provides a way of performing the computer-intensive study of the sampling behaviour of econometric estimators. Efron (1979, 1982a) proposed the bootstrap technique, which can be used to measure the variability of an estimator based on the available data. Freedman and Peters (1984a, 1984b) used bootstrapping to calculate the variability in the estimates of feasible GLS, two stage least squares and three stage least squares. Hinkley (1988) mentioned that the bootstrap method is in essence "the simulation of relevant properties of a statistical procedure with minimal model assumptions (p. 321)." In general, the distribution of a statistic

$S(X_1,\ldots,X_T)$ defined on a random sample $X_1,\ldots,X_T$ depends on the theoretical distribution F, which is unknown. Therefore, the nonparametric bootstrap approach uses the empirical distribution $\hat{F}$ instead of any assumed distribution. The rational choice of $\hat{F}$ is

$$\hat{F}(x) = \frac{1}{T}\sum h(x - x_i)$$

where

$$\begin{aligned} h(u) &= 0 \quad \text{if } u < 0 \\ &\quad 1 \quad \text{if } u \geq 0 \end{aligned}$$

Consider a general regression model as follows:

$$y_t = g(\underline{x}_t, \underline{\beta}) + \epsilon_t,\ t = 1,\ldots,T \tag{3.16}$$

where $g(\cdot)$ has a general functional form depending on the known data matrix, $\underline{\beta}$ is $K \times 1$ vector of unknown parameters, and ϵ_t is identically and independently distributed as F with $E_F(\epsilon_t) = 0$. Using vector notation,

$$\underline{y} = \underline{g}(X, \underline{\beta}) + \underline{\epsilon} \tag{3.17}$$

Then $\hat{\underline{\beta}}$ is estimated by minimizing some distance measure $D(\underline{y}, \underline{g})$, which is usually the square of Euclidean distance. The bootstrap provides a way of getting the covariance of the estimator $\hat{\underline{\beta}}$ by simulation rather than by using asymptotics. Therefore, this covariance matrix measures the finite sample variability of an estimator and can be quite useful when we know only the asymptotic properties of

the estimators. The bootstrap procedure in the context of the regression model is summarized as follows:

1. Assume $\hat{F}$ by $1/T$ at each estimated residual, $\hat{\epsilon}_t = y_t - g_t(\underline{\hat{\beta}})$.

2. Draw simulated random samples of size T with replacement from $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_T$. Calculate the bootstrap observations of y_t

$$y_t^* = g(\underline{x}_t, \underline{\hat{\beta}}) + \epsilon_t^*, \; t = 1, \ldots, T$$

3. Estimate $\underline{\hat{\beta}}^*$ which minimizes $D(\underline{y}^*, \underline{g}) = (\underline{y}^* - \underline{g})'(\underline{y}^* - \underline{g})$.

4. Repeat 2 and 3 many(B) times.

5. Calculate the bootstrap covariance matrix and bias of $\underline{\hat{\beta}}$

$$\begin{aligned} COV_*(\underline{\hat{\beta}}) &= \frac{1}{B-1}\sum_{b=1}^{B}(\underline{\hat{\beta}}^{*b} - E_*(\underline{\hat{\beta}}))(\underline{\hat{\beta}}^{*b} - E_*(\underline{\hat{\beta}}))' \\ Bias_*(\underline{\hat{\beta}}) &= E_*(\underline{\hat{\beta}}) - \underline{\hat{\beta}} \end{aligned}$$

where

$$E_*(\underline{\hat{\beta}}) = \frac{1}{B}\sum_{b=1}^{B}\underline{\hat{\beta}}^{*b}$$

We wish to compare the estimated standard errors of the maximum likelihood estimators with the bootstrap estimates of variability in the context of finite samples. Since nonlinear ML estimation relies on iterative procedures which require

much computational time and cost, we will limit the application of the bootstrap to one model ($\lambda_1 = 0.1$) for the cases $T = 30$ and $T = 60$.

It is rare that the exact sampling distribution of estimators and related test statistics are known. The usual statistical inference depends on large sample asymptotics. There have been a series of research efforts aimed at deriving the exact distributions for econometric estimators and test statistics. Phillips (1982) and Rothenberg (1982) give a good survey on the current state of the art in this field. In addition, Taylor's critique [(1983), p. 31] on finite sample distribution theory merits attention:

> *The most critical difficulties in applied econometrics remain model selection and specification, and nothing we have cited aids a whit in that process. Worse than that, our useful finite sample results seem to be computationally feasible only for moderately small models and classical conditions — precisely circumstances in which the adequacy of the model is most subject to criticism. It is an unfortunate irony that these techniques work best where they are needed least.*

The Edgeworth expansion is an important technique in studying the exact sampling distributions of estimators and test statistics. The popularity of this technique comes from the close relationship with the commonly used large sample theory and the dependence on the normal and chi-square distributions. Let $F_T(x) = F(x) + o(T^{-k})$ if $\lim_{T\to\infty} T^k|F_T - F| = 0$ for all x. Then the asymptotic

distribution of F_T is an $o(1)$ approximation to F. For example, the ML estimator $\hat{\beta}_T$ with distribution F_T has a limiting normal distribution, so that

$$F_T(x) = \Phi(x) + o(1),$$

where $\Phi(x)$ is a standard normal distribution function and the asymptotic variance of $\hat{\beta}_T$ is assumed to be 1. Suppose the random variables $X_1, \ldots, X_T$ each have the continuous probability density function f with mean zero, variance 1, and moments up to 4th order. Then the cumulant generating function has a power series expansion in a neighborhood of mean zero:

$$\ln \psi(t) = \frac{1}{2}(it)^2 + \frac{1}{6}\kappa_3(it)^3 + \frac{1}{24}\kappa_4(it)^4 + \cdots$$

where $\psi(t) = E(e^{itx})$ and κ_j = jth cumulant of f. The jth cumulant of the standardized sum $S_T = \frac{1}{\sqrt{T}} \sum X_t$ is $\gamma_j = \kappa_j T^{(1-j/2)}$, for $j > 0$. Therefore, higher order cumulants of S_T are close to zero as $T \to \infty$. Let $\psi_T(t)$ be the characteristic function of S_T. The cumulant generating function of S_T is

$$\ln \psi_T(t) = \frac{1}{2}(it)^2 + \frac{1}{6}\gamma_3(it)^3 + \frac{1}{24}\gamma_4(it)^4 + \cdots$$

Thus

$$\psi_T(t) = \exp(-t^2/2)[1 + \frac{1}{6}\gamma_3(it)^3 + \frac{1}{24}\gamma_4(it)^4 + \frac{1}{72}\gamma_3^2(it)^6 + \cdots]$$

Using the Fourier inversion formula, if $\int_{-\infty}^{+\infty} |\psi(t)|dt < \infty$,

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx}\psi(t)dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itx+\ln\psi(t)}dt$$

and we can obtain the formal Edgeworth expansion of the probability density function of S_T:

$$\begin{aligned} f_T(x) &\approx \phi(x) - \frac{1}{6}\gamma_3\phi^{(3)}(x) + \frac{1}{24}\gamma_4\phi^{(4)}(x) + \frac{1}{72}\gamma_3^2\phi^{(6)}(x) \\ &= \phi(x)[1 - \frac{1}{6}\gamma_3\frac{\phi^{(3)}(x)}{\phi(x)} + \frac{1}{24}\gamma_4\frac{\phi^{(4)}(x)}{\phi(x)} + \frac{1}{72}\gamma_3^2\frac{\phi^{(6)}(x)}{\phi(x)}] \end{aligned} \tag{3.18}$$

where $\phi^{(k)}(x)$ is the kth derivative of the standard normal density function $\phi(x)$. Integrating (3.18), we obtain the approximate distribution function of S_T:

$$F_T = \Phi(x) - \phi(x)[\frac{1}{6}\gamma_3\frac{\phi^{(2)}(x)}{\phi(x)} - \frac{1}{24}\gamma_4\frac{\phi^{(3)}(x)}{\phi(x)} - \frac{1}{72}\gamma_3^2\frac{\phi^{(5)}(x)}{\phi(x)}] \tag{3.19}$$

or alternatively

$$F_T = \Phi\{x - \frac{1}{6}\gamma_3(x^2-1) + \frac{1}{72}[3\gamma_4(3x - x^3) + 2\gamma_3^2(4x^3 - 7x)]\} \tag{3.20}$$

Equations (3.19) and (3.20) are the $o(T^{-1})$ Edgeworth expansion of $F_T(x)$.

The exact distribution of the t statistic has been studied within the framework of the simultaneous equations model [Richardson and Rhor (1971); Sargan (1975); Tse(1984)]. But an analytical study of the exact distributions of asymptotic t ratios in a nonlinear model has not yet been performed using the Edgeworth expansion. The ML estimators of the Box-Cox transformation cannot be represented by the data matrix only, so their behaviour is investigated using the limiting distribution. To get the Edgeworth expansion up to $o(T^{-1})$ of the t ratios of the Box-Cox model, we assume that

$$\sqrt{T}(\underline{\theta} - \underline{\theta}_0) \rightarrow N(\underline{0}, \Sigma)$$

where $\Sigma = (\sigma_{ij})$ and $\sigma_i^2 = \sigma_{ii}$. The conditions under which the density of $\sqrt{T}(\underline{\theta} - \underline{\theta}_0)$ has a valid Edgeworth expansion are discussed by Phillips (1977) and Sargan and Satchell (1986). The standardized variable $z_i = \sqrt{T}(\theta_i - \theta_{0i})/\sigma_i$ has the $N(0,1)$ distribution asymptotically. Define the t ratio:

$$\tau = \sqrt{T}(\theta_i - \theta_{0i})/\hat{\sigma}_i \tag{3.21}$$

where $\hat{\sigma}_i^2$ is a consistent estimator of σ_i^2 and τ^2 is a Wald statistic which converges in distribution to $\chi^2_{(1)}$. Let

$$\hat{s}_i = \hat{\sigma}_i^{-1}, \quad s_i^{(1)} = \frac{\partial s_i}{\partial \theta_i}, \quad s_i^{(2)} = \frac{\partial^2 s_i}{\partial \theta_i^2}$$

Develop the Taylor series expansion of $\hat{s}_i$:

$$\hat{s}_i = s_i + \frac{\sigma_i}{\sqrt{T}} s_i^{(1)} z_i + \frac{\sigma_i^2}{2T} s_i^{(2)} z_i^2 + O_P(T^{-3/2}) \tag{3.22}$$

Substitute Equation (3.22) into (3.21):

$$\tau = \sigma_i z_i \hat{s}_i = z_i + \frac{\sigma_i^2}{\sqrt{T}} s_i^{(1)} z_i^2 + \frac{\sigma_i^3}{2T} s_i^{(2)} z_i^3 + O_P(T^{-2/3})$$

The characteristic function for τ is

$$\begin{aligned}
\psi_\tau(t) &= (2\pi)^{-1/2} \int_{\mathcal{R}} \exp(it\tau - z_i^2/2) dz_i \\
&= (2\pi)^{-1/2} \int_{\mathcal{R}} \exp[-(z_i - it)^2/2 - t^2/2] \cdot \\
&\quad \{1 + [\frac{\sigma_i^2}{\sqrt{T}} s_i^{(1)} z_i^2 + \frac{\sigma_i^2}{2T} s_i^{(2)} z_i^3](it) + \frac{\sigma_i^4}{2T} s_i^{(1)^2} z_i^4 (it)^2\} dz_i \\
&\quad + o(T^{-1}) \\
&= e^{-t^2/2}\{1 + \frac{\sigma_i^2}{\sqrt{T}} s_i^{(1)}(it) + [\frac{3\sigma_i^3}{2T} s_i^{(2)} + \frac{3\sigma_i^4}{2T} s_i^{(1)^2}](it)^2
\end{aligned}$$

$$+\frac{\sigma_i^2}{\sqrt{T}}s_i^{(1)}(it)^3 + [\frac{\sigma_i^3}{2T}s_i^{(2)} + \frac{3\sigma_i^4}{T}s_i^{(1)^2}](it)^4$$
$$+\frac{\sigma_i^4}{2T}s_i^{(1)^2}(it)^6\} + o(T^{-1})$$

where i= $\sqrt{-1}$. Consider the relationship for Fourier inversion [Cramer(1946), p. 225]:

$$\int_{-\infty}^{+\infty} e^{itx}\phi^{(\nu)}(x)dx = (-it)^\nu e^{-t^2/2}, \quad \nu = 0,1,2,\cdots$$

The probability density function of τ is obtained from $\psi_\tau(t)$ as follows:

$$\begin{aligned} f_\tau(x) &= \phi(x) - \frac{\sigma_i^2}{\sqrt{T}}s_i^{(1)}\phi^{(1)}(x) + [\frac{3\sigma_i^3}{2T}s_i^{(2)} + \frac{3\sigma_i^4}{2T}s_i^{(1)^2}]\phi^{(2)}(x) \\ &\quad - \frac{\sigma_i^2}{\sqrt{T}}s_i^{(1)}\phi^{(3)}(x) + [\frac{\sigma_i^3}{2T}s_i^{(2)} + \frac{3\sigma_i^4}{T}s_i^{(1)^2}]\phi^{(4)}(x) \\ &\quad + \frac{\sigma_i^4}{2T}s_i^{(1)^2}\phi^{(6)}(x) + o(T^{-1}) \end{aligned}$$

Then we can obtain the distribution function by integrating $f_\tau(x)$:

$$\begin{aligned} F_\tau(x) &= \Phi(x) - \frac{\sigma_i^2}{\sqrt{T}}s_i^{(1)}\phi(x) + [\frac{3\sigma_i^3}{2T}s_i^{(2)} + \frac{3\sigma_i^4}{2T}s_i^{(1)^2}]\phi^{(1)}(x) \\ &\quad - \frac{\sigma_i^2}{\sqrt{T}}s_i^{(1)}\phi^{(2)}(x) + [\frac{\sigma_i^3}{2T}s_i^{(2)} + \frac{3\sigma_i^4}{T}s_i^{(1)^2}]\phi^{(3)}(x) \\ &\quad + \frac{\sigma_i^4}{2T}s_i^{(1)^2}\phi^{(5)}(x) + o(T^{-1}) \\ &= \Phi(x) - \phi(x)\{\frac{\sigma_i^2}{\sqrt{T}}s_i^{(1)} + [\frac{3\sigma_i^3}{2T}s_i^{(2)} + \frac{3\sigma_i^4}{2T}s_i^{(1)^2}]H_1(x) \\ &\quad + \frac{\sigma_i^2}{\sqrt{T}}s_i^{(1)}H_2(x) + [\frac{\sigma_i^3}{2T}s_i^{(2)} + \frac{3\sigma_i^4}{T}s_i^{(1)^2}]H_3(x) \\ &\quad + \frac{\sigma_i^4}{2T}s_i^{(1)^2}H_5(x)\} + o(T^{-1}) \end{aligned} \tag{3.23}$$

where $H_i(x)$ is a Hermite polynomial such that $H_i(x) = (-1)^i\phi^{(i)}(x)/\phi(x)$, or Equation (3.23) can be written as an explicit polynomial form:

$$F_\tau(x) = \Phi(x) - \phi(x)\{\frac{\sigma_i^2}{\sqrt{T}}s_i^{(1)} + [\frac{3\sigma_i^3}{2T}s_i^{(2)} + \frac{3\sigma_i^4}{2T}s_i^{(1)^2}]x$$

$$+\frac{\sigma_i^2}{\sqrt{T}}s_i^{(1)}(x^2-1)+[\frac{\sigma_i^3}{2T}s_i^{(2)}+\frac{3\sigma_i^4}{T}s_i^{(1)^2}](x^3-3x)$$
$$+\frac{\sigma_i^4}{2T}s_i^{(1)^2}(x^5-10x^3+15x)+o(T^{-1}) \qquad (3.24)$$

There is no correction at $x=0$ since $F_\tau(0)=\Phi(0)=0.5$. Suppose σ_i is a linear function of θ_i. Then $s_i^{(2)}=0$. Therefore, at $x=2$ Equation (3.24) can be written as

$$F_\tau(2)=\Phi(2)+\phi(2)\frac{4}{\sqrt{T}}\frac{\partial\sigma_i}{\partial\theta_i}$$

If $\frac{\partial\sigma_i}{\partial\theta_i}$ is negative, the corrected distribution has smaller probability than the standard normal distribution at 2. In general, the correction term is negligible for large values of T.

In the Box-Cox transformation it is difficult to calculate σ_i, $s_i^{(1)}$, and $s_i^{(2)}$. Therefore, the corrected distribution (3.23) is of no use in practice. Hall (1988) suggested a method for improving the test procedure by inverting a general Edgeworth expansion. Suppose the test statistic u_T with asymptotic mean zero and asymptotic variance one admits the following Edgeworth expansion:

$$P[u_T\le x]=\Phi(x)+T^{-1/2}\eta_{11}(x)\phi(x)+T^{-1}\eta_{12}(x)\phi(x)+\cdots$$

The inversion of the expansion up to $O(T^{-k/2})$ has the form:

$$P[u_T\le x-T^{-1/2}\hat{\eta}_1(x)-\cdots-T^{-(k-1)/2}\hat{\eta}_{k-1}]=\Phi(x)+O(T^{-k/2})$$

where $\hat{\eta}_i$ $(i=1,\cdots,k-1)$ is an estimator of η_i and the function $\hat{\eta}_i$ is assumed to be smooth in order to admit the Edgeworth expansion. But caution must be

used when using the inversion technique since overcorrection may occur when the sample size is small and the number of correction terms is large. In practice, it is difficult or even impossible to estimate the correction terms η_i.

The bootstrap inversion procedure can be used to cope with this problem [Rayner (1989)]. Suppose $Y_1, \cdots, Y_T$ are independent and identically distributed random variables. Then the bootstrap sample $Y_1^*, \cdots, Y_T^*$ is assumed to be iid $\hat{F}$ with mass $1/T$ at y_t. Define the bootstrap t ratio as

$$u_T^* = \sqrt{T}(\hat{\theta}^* - \theta)/\hat{\sigma}^*$$

where $\hat{\theta}^* = \hat{\theta}(Y_1^*, \cdots, Y_T^*)$ and $\hat{\sigma}^{2*} = \hat{\sigma}^2(Y_1^*, \cdots, Y_T^*)$. Therefore, the bootstrap distribution $P_T^* = P(u_T^* \leq x)$ has the following Edgeworth expansion:

$$P[u_T^* \leq x] = \Phi(x) + T^{-1/2}\hat{\eta}_{11}(x)\phi(x) + T^{-1}\hat{\eta}_{12}(x)\phi(x) + \cdots$$

Rayner (1989) suggested the following proposition supporting the bootstrap method of Edgeworth inversion:

Proposition 3.1 *Let u_T be a studentized statistic in a broad sense. Suppose sufficiently many finite moments of Y and derivatives of u_T exist so that u_T admits an Edgeworth expansion valid to order $O(T^{-(k+1)/2})$, i.e.,*

$$P[u_T \leq x] = \Phi(x) + \sum_{j=1}^{k+1} T^{-j/2}\phi(x)\eta_{1j} + O(T^{-(k+1)/2})$$

uniformly in x. Let

$$x_j = x_{j-1} - [\hat{P}_T^{*j}(x_{j-1}) - \Phi(x)]/\phi(x), \quad j = 1, \ldots, k$$

where $x_0 = x$ *and* $\hat{P}_T^{*j}(x) = \#\{u_T^{*j} < x_{j-1}\}/B_j$. *The notation* u_T^{*j} *denotes the t ratio obtained from the* j*th bootstrap for the correction of the confidence interval up to* $O(T^{-(k+1)/2})$. *Then*

$$P[u_T \leq x_k] = \Phi(x) + O(T^{-(k+1)/2})$$

uniformly in x *on compact intervals as* $T \to \infty$ *if*

$$B_j \geq T^{k+1+\delta} \ln(T) \quad \textit{for any } \delta > 0,\ j = 1, \ldots, k$$

Example:

For k=1,

$$P(u_T \leq x_1) = \Phi(x_\alpha) + O(T^{-1})$$

where

$$\begin{aligned} x_1 &= x_\alpha - [\hat{P}_T^{*1} - \Phi(x_\alpha)]/\phi(x_\alpha) \\ x_\alpha &= \Phi^{-1}(\alpha) \end{aligned}$$

For k=2,

$$P(u_T \leq x_2) = \Phi(x_\alpha) + O(T^{-3/2})$$

where

$$\begin{aligned} x_2 &= x_1 - [\hat{P}_T^{*2} - \Phi(x_\alpha)]/\phi(x_\alpha) \\ &= x_\alpha - [\hat{P}_T^{*1} - \Phi(x_\alpha)]/\phi(x_\alpha) - [\hat{P}_T^{*2} - \Phi(x_\alpha)]/\phi(x_\alpha) \end{aligned}$$

3.4 Design of Monte Carlo Simulation

Monte Carlo simulation is a widely used method for studying finite sample properties. Our concern is with the bias and precision of maximum likelihood estimators and the empirical distribution of their t ratios in samples of size 30 and 60. Extensive Monte Carlo simulation will be applied to the general transformation-of-variables model to investigate these properties.

For the purpose of analysis, let the Box-Cox model be

$$y_t^{(\lambda_1)} = \beta_1 + \beta_2 x_{2t}^{(\lambda_2)} + \beta_3 x_{3t}^{(\lambda_3)} + \epsilon_t \tag{3.25}$$

where $\beta_1 = \text{sign}(\lambda_1) * 10.0$, $\beta_2 = \text{sign}(\lambda_1) * 1.5$, $\beta_3 = -\text{sign}(\lambda_1) * 0.5$, $\lambda_2 = 0.1$, $\lambda_3 = 1.0$, and $\lambda_1 =$ (-1.0, -0.5, 0.1, 0.5, 1.0). The signs of the parameters (β_1, β_2 and β_3) must vary according to the sign of λ_1 in order for the right-hand side of (3.25) to be positive; $0 < y_t^{(\lambda_1)} < 1$ if $\lambda_1 < 0$; $y_t^{(\lambda_1)} > 1$ if $\lambda_1 > 0$. The above values of λ_1 are selected to contain the linear and nearly log-linear model since typical economic applications use linear and log-linear models. The model is chosen to have two explanatory variables whose values are generated from linear combinations of uniform random numbers so that $corr(x_{2t}, x_{3t}) = 0.2$:

$$\begin{aligned} x_{2t} &= 12 + 4U_{1t} \\ x_{3t} &= 15 + U_{1t} + 2\sqrt{6}U_{2t} \end{aligned}$$

where U_{1t} and $U_{2t} \sim U(-\sqrt{3}, \sqrt{3})$. Therefore, the probability density function of

x_{2t} and x_{3t} becomes

$$
\begin{aligned}
g_1(x_{2t}) &= \frac{1}{6\sqrt{2}} && \text{if } 5.0718 \leq x_{2t} \leq 18.9282 \\
&= 0 && \text{otherwise} \\
g_2(x_{3t}) &= (x_{3t} - 4.7287)/36 && \text{if } 4.7827 \leq x_{3t} < 8.2468 \\
&= \frac{1}{6\sqrt{3}} && \text{if } 8.2468 \leq x_{3t} < 21.7532 \\
&= (-x_{3t} + 25.2173)/36 && \text{if } 21.7352 \leq x_{3t} \leq 25.2173 \\
&= 0 && \text{otherwise}
\end{aligned}
$$

The error disturbances will be obtained using the random number generator RNDNS provided in the GAUSS matrix language (version 2.0). The variance of the error term ($\sigma^2 = 0.1$) is selected to account for 2 % of the variation of the right hand side of the Box-Cox model and to prevent the large truncation of the dependent variable from seriously violating the normality assumption. GAUSS is used to estimate the Box-Cox transfomation. Each Box-Cox model (λ_1 = -1.0, -0.5, 0.1, 0.5, 1.0) is estimated 1000 times for the samples of size $T = 30$ and $T = 60$. From the uniform random numbers, 500 bootstrap resamples will be generated.

3.5 Results

The termination of the iterative procedure for the maximum of the log-likelihood function was determined by two criteria — the difference between successive values of the log-likelihood and the rate of change in the parameter estimates. The

estimator value $\underline{\hat{\theta}}_i$ was accepted as the maximum likelihood estimate if

$$\max_{1 \leq k \leq m} \left| \frac{\theta_{i+1,k} - \theta_{i,k}}{\theta_{i,k} + 10^{-3}} \right| < 10^{-4}$$

and

$$|\ell_{i+1} - \ell_i| < 10^{-4}$$

where $m = dim(\underline{\theta})$. The starting values are chosen as follows:

1. Consider the modified model:

$$y_t^{(\lambda_1)} = \beta_1 + \beta_2 x_{2t}^{(\lambda_2^0)} + \beta_3 x_{3t}^{(\lambda_3^0)} + \beta_2(\lambda_2 - \lambda_2^0)\frac{\partial x_{2t}^{(\lambda_2)}}{\partial \lambda_2}|_{\lambda_2^0} + \beta_3(\lambda_3 - \lambda_3^0)\frac{\partial x_{3t}^{(\lambda_3)}}{\partial \lambda_3}|_{\lambda_3^0} + \epsilon_t \qquad (3.26)$$

   where λ_i^0 $(i = 2, 3)$ is the true value of λ_i. Under the linearity condition, $\lambda_i^0 = 1$ can be used since usually the true value of λ_i is not known.

2. Varying the value of λ_i, apply OLS to Equation (3.26).

3. Find $\underline{\hat{\theta}}_0$ which maximizes the log-likelihood function within the range of values of λ_i, i.e., $\lambda_i \in (-1, 1)$.

Proposition 3.2 *The estimator obtained from the grid search method applied to* Equation (3.26) *is consistent.*

Proof:

The consistency is straightforward from the first order conditions of ML estimation:

$$\begin{aligned} x_{it}^{(\hat{\lambda}_i)}\hat{\epsilon}_t &= 0, \quad i = 2,3 \quad and \quad t = 1,\ldots,T \\ \frac{\partial x_{it}^{(\hat{\lambda}_i)}}{\partial \lambda_i}\hat{\epsilon}_t &= 0, \quad i = 2,3 \quad and \quad t = 1,\ldots,T \end{aligned}$$

Both the Newton-Raphson and IGLS algorithms were convergent within 100 iterations while on average, two or three percent of the simulation replicates showed the near singularity of Hessian matrix at the convergent point.

3.5.1 Bias and Standard Error

The two algorithms (Newton-Raphson and IGLS) produced almost the same estimates and root mean square errors for $\hat{\underline{\theta}}$. In general, there is no dominating rule in terms of bias. The biases, reported in Table 3.1, of $\hat{\beta}_2$ and $\hat{\beta}_3$ are significantly different from zero in all cases and they are over-estimated when the sign of the true parameter is positive whereas they are under-estimated when the true parameter is negative. The power transformation estimator for the dependent variable is biased upward by a significant amount when its true value has a negative sign but is biased downward when the sign is positive, at T=30. In Table 3.2, mean absolute errors are indicative of the consistency of Box-Cox ML estimators since all values are reduced as the sample size becomes larger. For $\hat{\beta}_1$ and $\hat{\beta}_2$, mean absolute errors seem to be large relative to those of other estimators.

We make the following observations concerning the root mean square error

(RMSE) and standard errors:

1. Standard errors from the IGLS covariance matrix appear to be a slightly better measure of RMSE than those from the inverted negative Hessian, except for β_2 and λ_2.

2. Generally, standard errors from the inverted negative Hessian and the IGLS covariance matrix are fairly close. For T=30, the standard errors of the linear parameter estimates are a poor approximation to RMSE but they become a better approximation to RMSE for T=60, especially for $\hat{\beta}_1$. Previously, standard errors of power parameter estimates were thought to be good approximations to RMSE's.

3. All standard errors and RMSE's become smaller as sample size increases.

4. The discrepancy between RMSE and standard error tends to disappear with increasing sample size. This observation is in agreement with the decreasing absolute errors as the sample size becomes larger.

5. At T=30, the standard errors of $\hat{\beta}_2$ are very small relative to RMSE while the bias estimates of $\hat{\beta}_2$ are large for all models (Table 3.1). We can infer that the ML method leads to a poor estimate of the linear parameter that corresponds to the explanatory variable with the nonlinear power transformation — in our case, $\lambda_2 = 0.1$.

Freedman and Peters (1984a) found that the standard errors for feasible GLS

regression coefficients are very small relative to the bootstrap standard deviations while the (nominal) standard errors of 3SLS estimators performed well in finite samples in their another study (1984b).

We chose one model ($\lambda_1 = 0.1$) in order to examine the behaviour of the nominal standard errors of the ML estimators of the Box-Cox model using bootstrapping. Bootstrapping was carried out with 500 replications for each of 1000 simulation estimates. The nominal standard errors were obtained by use of IGLS. In Table 3.5, we can observe that about 75 percent of the nominal standard errors fall within the range of 0.8–0.9 of the variability (bootstrap standard deviation) of the estimates obtained via bootstrap at T=30. When the sample size increases, standard errors are close to the bootstrap finite sample variability. In general, the nominal standard errors of ML estimators in the Box-Cox model are not seriously under-estimated according to our bootstrap experiment. But the bootstrap standard deviations are too small relative to the true sample variability — i.e., RMSE — at T=30. The difference between them is small when the sample size becomes larger, except for $\hat{\beta}_2$. Therefore, the bootstrap standard deviation is thought to be a poor indicator of finite sample variability when the sample size is relatively small.

3.5.2 Empirical Distribution

Empirical researchers usually depend on the asymptotic t ratio $(\hat{\theta}_i - \theta_i)/\hat{\sigma}_i$ in determining the significance of the parameters. But the standard normal distri-

bution may not provide a good approximation to the t ratio in small samples. The finite sample distribution was compared with the standard normal distribution using the Kolmogorov-Smironov test statistic. In Table 3.6, the results give evidence against the normality assumption. In particular, the t ratios for $\hat{\beta}_2$ and $\hat{\beta}_3$ show considerable distance from the standard normal distribution at $T = 30$. For these parameters, the difference between the sampling distribution and the normal distribution becomes smaller as sample size increases though the distance is sizable.

Deviations from Nominal Size in Linear Parameters

To study the empirical distribution of the t ratios, the deviations of real size from nominal size were calculated. The Empirical (real) size of the test statistics was obtained using the formula:

$$\#\{|t_i| > x_\alpha\}/N$$

where x_α denotes $\Phi^{-1}(1-\alpha/2)$ and N the number of replicates of simulation. The deviations from nominal size ($\alpha = 0.05$ and $\alpha = 0.10$) at $T = 30$ and $T = 60$ are tabulated in Tables 3.7–3.10. Under the null hypothesis $H_0 : \underline{\theta} = \underline{\theta}_0$, the deviations from nominal size in $\hat{\beta}_2$ and $\hat{\beta}_3$ are relatively large and positive at $T = 30$ for both $\alpha = 0.05$ and $\alpha = 0.10$. However, deviations are greatly reduced as the sample size increases. Deviations are larger at $\alpha = 0.05$ than at $\alpha = 0.10$ for $T = 30$ and $T = 60$. This may indicate that tail areas of the t ratios for $\hat{\beta}_2$ and $\hat{\beta}_3$ are relatively

heavy. The deviations for $\hat{\beta}_1$ are small relative to those of $\hat{\beta}_2$ and $\hat{\beta}_3$. When the hypothesized value is slightly greater than the true value, deviations decrease with the positive sign of the true linear parameter and deviations increase with the negative sign of true linear parameters. The reverse is true when the hypothesized value is less than the true value for all nominal sizes and sample sizes. Therefore, the t ratios of the linear parameters appear to be biased.

Deviations from Nominal Sizes in Power Parameters

Under the null hypothesis $H_0 : \underline{\theta} = \underline{\theta}_0$, the deviations in power parameters are shown to be slightly larger at $\alpha = 0.10$ than at $\alpha = 0.05$ for $T = 30$. When the null hypothesis is not true, the deviations are larger than in the case of a true null hypothesis. We observe that the t ratios of the power parameter estimates are unbiased since the rejection rate of the null hypothesis is smaller under the true null hypothesis than under the false null hypothesis.

3.5.3 Bootstrap Inversion of Edgeworth Expansion

Since the deviations from nominal size in the t ratios of $\hat{\beta}_2$ and $\hat{\beta}_3$ are conspicuous and their Kolmogorov-Smironov test statistics indicate a remarkable distance from a standard normal distribution, we made a bootstrap inversion of the Edgeworth expansion [Rayner (1989)] with only the first stage adjustment in order to improve the confidence interval. We limited the number of bootstrap resamples to B=300. Results are shown in Table 3.11. The Edgeworth expansion yields a

good approximation to the distribution of $\hat{\beta}_2$ and $\hat{\beta}_3$ except the case of $\alpha = 0.10$ at $T = 60$. When $\alpha = 0.05$, empirical confidence intervals become closer to 0.95 with size correction than with the original nominal size. The Edgeworth expansion also leads to better approximations of the confidence interval 0.90 when $\alpha = 0.10$. We observe that the Edgeworth size correction performs better at $T = 60$ than at $T = 30$, but overcorrection occurs for $\alpha = 0.10$ at $T = 60$.

The mean values of first stage adjustments and their standard deviations are also given in Table 3.11. The standard deviations of these adjustments are relatively large since the adjustment term is a function of higher moments which have large variability in small samples.

3.6 Conclusions

There are several important implications of this analysis of general transformation of variables model. We can summarize the conclusions as follows:

1. There was no difference between Newton-Raphson and IGLS algorithms since they yielded identical parameter estimates and similar variance estimates. The choice between these two algorithms can be made with the consideration of computational convenience.

2. The biases of the linear parameter estimates which correspond to the variable with the power transformation were sizable and significantly different from zero in small samples. But the estimators of the general Box-Cox model

parameters were shown to be consistent.

3. Standard errors from Newton-Raphson and IGLS under-estimated the finite sample variability at T=30. The standard errors of the power parameters, however, are close to RMSE. The standard error of $\hat{\beta}_2$ was substantially under-estimated at both sample sizes.

4. When we estimate the Box-Cox model via the usual ML method, the estimate of the linear parameter that corresponds to the variable with the nonlinear power transformation and its standard error seem to be unreliable.

5. In our experiment, the nominal standard errors were not too small compared to the bootstrap standard deviation. But the bootstrap finite sample variability turned out to be a poor approximation to the true finite sample variability (RMSE) in small samples (T=30).

6. There were biases in the t ratios for the linear parameters, while those for the power parameters did not show any bias. The t ratios of the linear parameters which correspond to the variable with power transformation were a bad approximation to the standard normal distribution. However, the Edgeworth correction via bootstrapping provided a better approximation to the finite sample distribution of these t ratios.

Table 3.1: Bias of MLE

	β_1	β_2	β_3	λ_1	λ_2	λ_3
$\lambda_1 = -1.0$						
T=30	-0.123	-1.445*	0.288*	0.023*	0.019	0.024*
	(0.220)	(0.155)	(0.036)	(0.007)	(0.016)	(0.008)
T=60	-0.425	-0.531*	0.146*	0.008	0.031*	0.002
	(0.155)	(0.070)	(0.017)	(0.005)	(0.012)	(0.005)
$\lambda_1 = -0.5$						
T=30	0.219	-2.006*	0.321*	0.009*	0.004	0.019*
	(0.234)	(0.229)	(0.040)	(0.004)	(0.017)	(0.008)
T=60	-0.065	-0.698*	0.134*	0.004	-0.004	0.002
	(0.142)	(0.074)	(0.017)	(0.003)	(0.012)	(0.005)
$\lambda_1 = 0.1$						
T=30	-0.331	1.825*	-0.271*	0.000	-0.008	0.003
	(0.177)	(0.196)	(0.028)	(0.001)	(0.017)	(0.008)
T=60	-0.015	0.783*	-0.027*	0.002	-0.020	-0.013*
	(0.116)	(0.077)	(0.013)	(0.001)	(0.012)	(0.005)
$\lambda_1 = 0.5$						
T=30	0.021	1.461*	-0.264*	0.011*	0.036*	0.030*
	(0.210)	(0.191)	(0.034)	(0.004)	(0.017)	(0.008)
T=60	-0.166	0.587*	-0.088*	-0.012*	0.000	0.015*
	(0.129)	(0.068)	(0.015)	(0.003)	(0.012)	(0.005)
$\lambda_1 = 1.0$						
T=30	0.006	1.725*	-0.305*	-0.027*	0.012	0.021*
	(0.242)	(0.190)	(0.039)	(0.007)	(0.016)	(0.008)
T=60	-0.123	0.778*	-0.120*	-0.021*	-0.039*	0.014*
	(0.155)	(0.070)	(0.017)	(0.005)	(0.011)	(0.005)

1) Bias $= \frac{\sum_i \hat{\theta}_{ij}}{N} - \theta_i$
2) Numbers in parenthesis are standard deviations of bias
3) Statistical significance at 5 % level is represented by *

Table 3.2: Mean Absolute Error of MLE

	β_1	β_2	β_3	λ_1	λ_2	λ_3
$\lambda_1 = -1.0$						
T=30	4.703	2.313	0.557	0.180	0.402	0.198
T=60	3.503	1.287	0.353	0.140	0.302	0.132
$\lambda_1 = -0.5$						
T=30	4.705	2.875	0.590	0.110	0.429	0.214
T=60	3.273	1.368	0.335	0.081	0.307	0.136
$\lambda_1 = 0.1$						
T=30	3.806	2.613	0.502	0.044	0.412	0.205
T=60	2.690	1.383	0.282	0.031	0.295	0.131
$\lambda_1 = 0.5$						
T=30	4.465	2.333	0.533	0.112	0.413	0.207
T=60	3.030	1.262	0.299	0.079	0.292	0.130
$\lambda_1 = 1.0$						
T=30	4.947	2.618	0.579	0.187	0.419	0.200
T=60	3.646	1.411	0.351	0.148	0.299	0.136

Mean Absolute Error $= \sum_{j=1}^{N} |\hat{\theta}_{ij} - \theta_i| / N$

where N = simulation replications (1000)

Table 3.3: RMSE and Standard Error of MLE (T=30)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
$\lambda_1 = -1.0$						
RMSE	6.878	5.061	1.156	0.230	0.512	0.253
SE_1/RMSE	0.87	0.66	0.65	0.93	0.95	0.89
SE_2/RMSE	0.92	0.66	0.71	0.98	0.88	0.92
$\lambda_1 = -0.5$						
RMSE	7.247	7.372	1.298	0.137	0.544	0.265
SE_1/RMSE	0.80	0.55	0.58	0.91	0.89	0.98
SE_2/RMSE	0.83	0.54	0.61	0.95	0.85	0.90
$\lambda_1 = 0.1$						
RMSE	5.463	6.294	0.909	0.055	0.524	0.262
SE_1/RMSE	0.86	0.58	0.69	0.89	0.92	0.89
SE_2/RMSE	0.89	0.57	0.73	0.94	0.87	0.92
$\lambda_1 = 0.5$						
RMSE	6.497	6.080	1.095	0.143	0.527	0.262
SE_1/RMSE	0.81	0.54	0.62	0.85	0.90	0.87
SE_2/RMSE	0.86	0.53	0.67	0.90	0.86	0.90
$\lambda_1 = 1.0$						
RMSE	7.585	6.199	1.269	0.239	0.524	0.254
SE_1/RMSE	0.81	0.58	0.62	0.89	0.90	0.89
SE_2/RMSE	0.86	0.52	0.66	0.94	0.87	0.93

$$\text{RMSE} = \sqrt{\sum_{j=1}^{N}(\hat{\theta}_{ij} - \theta_i)^2/N}$$

SE_1 = standard error from the Hessian matrix
SE_2 = standard error from IGLS covariance matrix

Table 3.4: RMSE and Standard Error of MLE (T=60)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
$\lambda_1 = -1.0$						
RMSE	4.889	2.272	0.558	0.178	0.377	0.166
SE_1/RMSE	0.89	0.78	0.78	0.93	1.03	0.91
SE_2/RMSE	0.99	0.75	0.90	1.04	0.96	1.01
$\lambda_1 = -0.5$						
RMSE	4.395	2.413	0.549	0.102	0.387	0.173
SE_1/RMSE	0.88	0.74	0.74	0.94	0.95	0.91
SE_2/RMSE	0.96	0.74	0.83	1.03	0.93	0.97
$\lambda_1 = 0.1$						
RMSE	3.550	2.476	0.416	0.039	0.373	0.164
SE_1/RMSE	0.93	0.73	0.90	0.98	0.98	1.00
SE_2/RMSE	0.99	0.74	0.96	1.05	0.96	1.03
$\lambda_1 = 0.5$						
RMSE	4.008	2.193	0.477	0.100	0.375	0.165
SE_1/RMSE	0.94	0.79	0.81	0.97	1.00	0.96
SE_2/RMSE	1.02	0.78	0.90	1.06	0.96	1.02
$\lambda_1 = 1.0$						
RMSE	4.875	2.358	0.566	0.187	0.373	0.170
SE_1/RMSE	0.87	0.83	0.74	0.89	1.02	0.90
SE_2/RMSE	0.96	0.80	0.85	0.99	0.96	0.98

$$\text{RMSE} = \sqrt{\sum_{j=1}^{N}(\hat{\theta}_{ij} - \theta_i)^2/N}$$

SE_1 = standard error from the Hessian matrix

SE_2 = standard error from IGLS covariance matrix

Table 3.5: Bootstrap Result for Model $\lambda_1 = 0.1$

	β_1	β_2	β_3	λ_1	λ_2	λ_3
T=30	%					
$0.9 < R_i \leq 1.0$	24.6	25.5	24.9	26.1	26.0	24.7
$0.8 < R_i \leq 0.9$	75.4	74.2	75.0	73.8	73.6	75.2
RMSE_i	5.463	6.294	0.909	0.055	0.524	0.262
$\bar{SD}_i$	4.926	3.634	0.676	0.053	0.467	0.248
$\bar{SE}_i$	4.339	3.206	0.595	0.046	0.412	0.218
$\bar{Bias}_i$	-0.353	1.744	-0.261	-0.001	-0.010	0.008
T=60	%					
$0.9 < R_i \leq 1.0$	88.2	88.0	88.7	89.7	88.2	88.3
$0.8 < R_i \leq 0.9$	8.3	8.8	7.6	7.1	8.4	8.3
RMSE_i	3.550	2.476	0.416	0.039	0.373	0.164
$\bar{SD}_i$	3.478	1.787	0.393	0.041	0.362	0.172
$\bar{SE}_i$	3.271	1.687	0.369	0.039	0.341	0.161
$\bar{Bias}_i$	-0.056	0.689	-0.107	0.000	0.000	-0.002

$\text{R}_i = SE_{ij}/SD_{ij}, \quad i = 1,\ldots,6, j = 1,\ldots,1000$

where $SE_{ij} = \sqrt{\frac{1}{B}\sum_{b=1}^{B} V(\hat{\theta}_{ij}^{*b})}$

$SD_{ij} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_{ij}^{*b} - E_*\hat{\theta}_{ij})^2}$

$V(\hat{\theta}_{ij}^{*b})$ = nominal variance estimator

$E_*(\hat{\theta}_{ij}) = \frac{\sum_{b=1}^{B}\hat{\theta}_{ij}^{*b}}{B}$

B = number of bootstrap replications (500)

$\text{RMSE}_i = \sqrt{\sum_{j=1}^{N}(\hat{\theta}_{ij} - \theta_i)^2/N}$

$\bar{SD}_i = \sum_j SD_{ij}/N$

$\bar{SE}_i = \sum_j SE_{ij}/N$

$\bar{Bias}_i = \sum_j \frac{E_*\hat{\theta}_{ij}}{N} - \theta_i$

Table 3.6: Kolmogorov-Smironov Statistics for Sampling Distribution

	β_1	β_2	β_3	λ_1	λ_2	λ_3
$\lambda_1 = -1.0$						
T=30	0.115	0.173	0.167	0.048	0.031	0.059
T=60	0.084	0.121	0.118	0.037	0.037	0.042
$\lambda_1 = -0.5$						
T=30	0.138	0.170	0.183	0.060	0.037	0.080
T=60	0.091	0.103	0.110	0.033	0.038	0.042
$\lambda_1 = 0.1$						
T=30	0.080	0.143	0.135	0.045	0.047	0.040
T=60	0.045	0.093	0.076	0.038	0.043	0.064
$\lambda_1 = 0.5$						
T=30	0.118	0.172	0.172	0.059	0.038	0.078
T=60	0.093	0.108	0.118	0.054	0.034	0.044
$\lambda_1 = 1.0$						
T=30	0.135	0.183	0.170	0.066	0.050	0.054
T=60	0.123	0.111	0.146	0.064	0.070	0.048

1) Critical values at level α are approximated by

$$c = [\frac{\ln(2/\alpha)}{2N}]^{1/2}$$

2) $P(D_N > c) = \alpha$

where α = significance level

$D_N = \sup_x |F_N(x) - \Phi(x)|$

$\Phi(x)$ = cdf of $N(0,1)$

F_N = sampling distribution

Table 3.7: Deviations from Nominal Size of N(0,1)
($\alpha = 0.05$, $T = 30$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
$\lambda_1 = -1.0$						
$\theta_H = \theta_0$	0.06	0.14	0.14	0.03	0.03	0.05
$\theta_H = \theta_0 - 0.5$	0.08	0.20	-0.05	0.63	0.16	0.57
$\theta_H = \theta_0 + 0.5$	0.04	0.07	0.33	0.56	0.17	0.51
$\lambda_1 = -0.5$						
$\theta_H = \theta_0$	0.03	0.14	0.15	0.04	0.03	0.04
$\theta_H = \theta_0 - 0.5$	0.06	0.20	-0.04	0.92	0.17	0.51
$\theta_H = \theta_0 + 0.5$	0.02	0.07	0.34	0.89	0.18	0.46
$\lambda_1 = 0.1$						
$\theta_H = \theta_0$	0.01	0.10	0.10	0.05	0.04	0.06
$\theta_H = \theta_0 - 0.5$	0.00	0.05	0.32	0.95	0.16	0.51
$\theta_H = \theta_0 + 0.5$	0.03	0.17	-0.03	0.95	0.19	0.50
$\lambda_1 = 0.5$						
$\theta_H = \theta_0$	0.05	0.14	0.14	0.05	0.03	0.06
$\theta_H = \theta_0 - 0.5$	0.04	0.07	0.34	0.89	0.17	0.55
$\theta_H = \theta_0 + 0.5$	0.07	0.20	-0.04	0.91	0.16	0.49
$\lambda_1 = 1.0$						
$\theta_H = \theta_0$	0.06	0.15	0.14	0.05	0.04	0.06
$\theta_H = \theta_0 - 0.5$	0.04	0.07	0.34	0.54	0.18	0.55
$\theta_H = \theta_0 + 0.5$	0.08	0.22	-0.04	0.63	0.19	0.49

1) Deviation $= P(|t_i| > x_\alpha) - 2 * (1 - \Phi(x_\alpha))$
where α = significance level
$x_\alpha = \Phi^{-1}(1 - \alpha/2)$
$\Phi(x)$ = cdf of $N(0,1)$
$t_i = (\hat{\theta}_i - \theta_i)/\hat{\sigma}_i$
$\hat{\sigma}_i$ = standard error from the Hessian

Table 3.8: Deviations from Nominal Size of N(0,1)
($\alpha = 0.05$, $T = 60$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
$\lambda_1 = -1.0$						
$\theta_H = \theta_0$	0.03	0.08	0.08	0.01	0.01	0.03
$\theta_H = \theta_0 - 0.5$	0.05	0.16	-0.02	0.81	0.20	0.86
$\theta_H = \theta_0 + 0.5$	0.01	0.10	0.36	0.74	0.20	0.81
$\lambda_1 = -0.5$						
$\theta_H = \theta_0$	0.03	0.06	0.08	0.02	0.02	0.04
$\theta_H = \theta_0 - 0.5$	0.04	0.13	-0.02	0.94	0.20	0.84
$\theta_H = \theta_0 + 0.5$	0.01	0.06	0.37	0.96	0.26	0.78
$\lambda_1 = 0.1$						
$\theta_H = \theta_0$	0.00	0.05	0.04	0.01	0.02	0.02
$\theta_H = \theta_0 - 0.5$	-0.01	0.00	0.33	0.95	0.20	0.81
$\theta_H = \theta_0 + 0.5$	0.01	0.13	0.05	0.95	0.27	0.79
$\lambda_1 = 0.5$						
$\theta_H = \theta_0$	0.02	0.07	0.08	0.00	0.02	0.01
$\theta_H = \theta_0 - 0.5$	0.00	0.01	0.40	0.94	0.19	0.85
$\theta_H = \theta_0 + 0.5$	0.04	0.15	-0.02	0.95	0.24	0.77
$\lambda_1 = 1.0$						
$\theta_H = \theta_0$	0.06	0.04	0.11	0.03	0.00	0.04
$\theta_H = \theta_0 - 0.5$	0.04	0.00	0.39	0.72	0.15	0.86
$\theta_H = \theta_0 + 0.5$	0.09	0.13	-0.03	0.80	0.27	0.78

1) Deviation $= P(|t_i| > x_\alpha) - 2 * (1 - \Phi(x_\alpha))$
where α = significance level
$\mathrm{x}_\alpha = \Phi^{-1}(1 - \alpha/2)$
$\Phi(x)$ = cdf of $N(0,1)$
$\mathrm{t}_i = (\hat{\theta}_i - \theta_i)/\hat{\sigma}_i$
$\hat{\sigma}_i$ = standard error from the Hessian

Table 3.9: Deviations from Nominal Size of N(0,1)
($\alpha = 0.10$, $T = 30$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
$\lambda_1 = -1.0$						
$\theta_H = \theta_0$	0.04	0.11	0.11	0.03	0.04	0.07
$\theta_H = \theta_0 - 0.5$	0.07	0.18	-0.08	0.66	0.21	0.62
$\theta_H = \theta_0 + 0.5$	0.03	0.04	0.32	0.62	0.20	0.57
$\lambda_1 = -0.5$						
$\theta_H = \theta_0$	0.04	0.11	0.12	0.05	0.05	0.07
$\theta_H = \theta_0 - 0.5$	0.07	0.19	-0.07	0.87	0.22	0.58
$\theta_H = \theta_0 + 0.5$	0.01	0.04	0.32	0.87	0.21	0.53
$\lambda_1 = 0.1$						
$\theta_H = \theta_0$	0.01	0.08	0.08	0.06	0.05	0.06
$\theta_H = \theta_0 - 0.5$	-0.01	0.01	0.31	0.90	0.19	0.57
$\theta_H = \theta_0 + 0.5$	0.04	0.15	-0.02	0.90	0.20	0.56
$\lambda_1 = 0.5$						
$\theta_H = \theta_0$	0.05	0.11	0.11	0.07	0.04	0.07
$\theta_H = \theta_0 - 0.5$	0.03	0.04	0.34	0.86	0.21	0.61
$\theta_H = \theta_0 + 0.5$	0.07	0.18	-0.07	0.88	0.21	0.55
$\lambda_1 = 1.0$						
$\theta_H = \theta_0$	0.05	0.12	0.12	0.05	0.05	0.06
$\theta_H = \theta_0 - 0.5$	0.02	0.04	0.33	0.60	0.22	0.61
$\theta_H = \theta_0 + 0.5$	0.08	0.20	-0.08	0.67	0.22	0.56

1) Deviation = $P(|t_i| > x_\alpha) - 2 * (1 - \Phi(x_\alpha))$
where α = significance level
$x_\alpha = \Phi^{-1}(1 - \alpha/2)$
$\Phi(x)$ = cdf of $N(0,1)$
$t_i = (\hat{\theta}_i - \theta_i)/\hat{\sigma}_i$
$\hat{\sigma}_i$ = standard error from the Hessian

Table 3.10: Deviations from Nominal Size of N(0,1)
($\alpha = 0.10$, $T = 60$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
$\lambda_1 = -1.0$						
$\theta_H = \theta_0$	0.03	0.06	0.06	0.02	0.00	0.05
$\theta_H = \theta_0 - 0.5$	0.06	0.15	0.09	0.82	0.27	0.85
$\theta_H = \theta_0 + 0.5$	0.01	0.02	0.36	0.76	0.24	0.81
$\lambda_1 = -0.5$						
$\theta_H = \theta_0$	0.02	0.05	0.05	0.02	0.03	0.05
$\theta_H = \theta_0 - 0.5$	0.05	0.12	0.14	0.89	0.26	0.83
$\theta_H = \theta_0 + 0.5$	0.01	-0.02	0.37	0.89	0.30	0.79
$\lambda_1 = 0.1$						
$\theta_H = \theta_0$	0.00	0.02	0.01	0.01	0.02	0.00
$\theta_H = \theta_0 - 0.5$	0.00	-0.03	0.34	0.90	0.26	0.81
$\theta_H = \theta_0 + 0.5$	0.01	0.11	0.43	0.90	0.32	0.80
$\lambda_1 = 0.5$						
$\theta_H = \theta_0$	0.01	0.05	0.06	0.02	0.00	0.02
$\theta_H = \theta_0 - 0.5$	0.00	-0.02	0.39	0.89	0.24	0.83
$\theta_H = \theta_0 + 0.5$	0.05	0.14	0.13	0.90	0.30	0.79
$\lambda_1 = 1.0$						
$\theta_H = \theta_0$	0.06	0.03	0.09	0.04	0.01	0.05
$\theta_H = \theta_0 - 0.5$	0.04	-0.05	0.37	0.75	0.21	0.85
$\theta_H = \theta_0 + 0.5$	0.09	0.12	-0.08	0.81	0.33	0.79

1) Deviation $= P(|t_i| > x_\alpha) - 2 * (1 - \Phi(x_\alpha))$
where α = significance level
$x_\alpha = \Phi^{-1}(1 - \alpha/2)$
$\Phi(x)$ = cdf of $N(0,1)$
$t_i = (\hat{\theta}_i - \theta_i)/\hat{\sigma}_i$
$\hat{\sigma}_i$ = standard error from the Hessian

Table 3.11: Bootstrap Inversion of Edgeworth Expansions

	β_2		β_3	
	T=30	T=60	T=30	T=60
$\alpha = 0.05$				
x_α	0.841	0.880	0.857	0.899
x_1	0.891	0.923	0.912	0.948
Mean C	-1.148	-0.737	-1.135	-0.735
SD_C	0.276	0.252	0.287	0.265
$\alpha = 0.10$				
x_α	0.813	0.849	0.832	0.873
x_1	0.871	0.899	0.892	0.930
Mean C	-1.009	-0.717	-0.997	-0.719
SD_C	0.197	0.185	0.198	0.178

1) $x_\alpha = \Phi^{-1}(1 - \alpha/2)$
2) $x_1 = x_\alpha - C$
3) C = $[\hat{P}_T^*(x_\alpha) - \Phi(x_\alpha)]/\phi(x_\alpha)$
where $\Phi(\cdot)$ = cdf of $N(0,1)$
$\phi(\cdot)$ = pdf of $N(0,1)$

CHAPTER 4
TESTING THE BOX-COX MODEL IN SMALL SAMPLES

4.1 Introduction

The linear and log-linear regression models are frequently used when specifying a functional form since they are simple and appealing in an economic sense. Godfrey and Wickens (1981) discussed procedures for testing the linear and log-linear models against the Box-Cox regression model. Using Box-Cox alternatives, they extended the traditional approach of testing linearity and log-linearity against each other. They considered two different test statistics, a Lagrange multiplier (LM) test (outer product form) and Andrews' (1971) exact test.

Davidson and Mackinnon (1983) studied the small sample properties of two variants of the LM test — outer products of gradient (OPG) and double length regression (DLR). In small samples, the OPG variant rejects the null hypothesis more often than the DLR variant. They concluded that the OPG form of the LM statistic might lead to serious errors of statistical inference. In the context of the Box-Cox regression model, Davidson and Mackinnon (1985) investigated the power properties of the two variants of the LM test, Andrews exact test and the transformed version of the likelihood ratio (LR) test. Their Monte Carlo experiment revealed that the classical tests (the OPG and DLR variants, and LR) were much more powerful than the Andrews test, and that the Andrews test

seriously lacks in power if the variance of the error term in the regression equation is large. On the whole, the power of the OPG variant was less than that of DLR and LR.

Draper and Cox (1969) suggested that the Box-Cox transformation could help regularize data and the estimated transformation might yield a nearly symmetric distribution. In addition, they obtained the variance of the power transformation estimator, which turned out to be incorrect since they used a non-regular likelihood function [Hinkley (1975); Amemiya and Powell (1981)].

We consider the following Box-Cox regression model:

$$\begin{aligned} y_t^{(\lambda_1)} &= \beta_1 + \beta_2 x_{2t}^{(\lambda_2)} + \cdots + \beta_k x_{kt}^{(\lambda_k)} + \epsilon_t \\ &= \mu_t + \epsilon_t, \quad t = 1, \ldots, T \end{aligned}$$

where ϵ_t has the symmetric distribution function F, $E_F(\epsilon_t) = 0$ and $E_F(\epsilon_t^2) = \sigma^2$. Let $\theta_t = sign(\lambda_1)\frac{\sigma}{1/\lambda_1 + \mu_t}$ which is the coefficient of variation of $y_t^{\lambda_1}$. If $1 + \lambda_1\mu_t$ is much greater than $|\lambda_1|\sigma$, i.e. $\theta_t << 1$, then each observation of the dependent variables y_t has very high probability of being positive and thus the small θ_t-approximation leads to a very good approximation of the asymptotic results [Draper and Cox (1969); Taylor (1985)]. Zarembka (1974) also used the small θ_t-approximation. He proposed that the Box-Cox procedure might yield an approximately consistent estimator of the power transformation parameter when the error term was reasonably symmetric and homoskedastic. He showed that maxi-

mum likelihood (ML) estimation was not robust to the heteroskedasticity of the error term.

The Box-Cox approach implies that $y_t^{(\lambda_1)}$ is not normally distributed unless $\lambda_1 = 0$. Therefore, the usual ML estimators based on the assumption of normality of the transformed dependent variable are not consistent and the asymptotic covariance matrix of the Box-Cox ML estimators is not the inverse of the information matrix [Hinkley (1975); Amemiya and Powell (1981)]. Therefore, it is necessary to assume a proper distribution function for the pre-transformed dependent variable y_t in the theoretical analysis of the Box-Cox transformation. Poirier (1978) used a truncated normal distribution; Amemiya and Powell (1981) used a gamma distribution.

When we estimate the model, the usual practice is to test individual coefficients using t-ratios. In nonlinear regression models, the asymptotic covariance matrix can be used to compute a t statistic. Recent studies [Griffiths, Hill and Pope (1987); Calzolari and Panattoni (1988)] have shown that the asymptotically equivalent covariance matrix estimators (or standard errors) are different in finite samples. Within the framework of the Box-Cox model, the functional forms of the transformed variables are tested using one of the asymptotically equivalent test statistics (LM, LR and Wald). The asymptotic equivalence of these statistics is not guaranteed in small samples. Furthermore, Nelson and Savin (1988) investigated the power functions of the LM, LR and Wald tests in one-parameter

nonlinear models and displayed the nonmonotonic power function of the Wald statistic. Therefore, the choice of the test procedure does not seem to be a matter of convenience or personal preference, especially in small samples.

The purpose of this Chapter is to investigate the small sample properties of the asymptotic standard errors of Box-Cox ML estimators. The empirical distribution and size of three asymptotically equivalent test statistics (LM, LR and Wald) for testing the functional form will also be examined in the context of small samples ($T = 30$ and $T = 60$). In addition, the estimated power functions of these asymptotic tests will be compared to the asymptotic power functions.

In Section 4.2, we wish to show the improper distribution characteristic of the dependent variable when we use the usual ML estimation method for the Box-Cox transformation. Asymptotic covariance matrix estimators and test statistics are discussed in Section 4.3. In Section 4.4, the design of a Monte Carlo simulation is given and in Section 4.5 the results are presented. Section 4.6 contains a summary and conclusions.

4.2 The Box-Cox Transformation: Model and Assumptions

Consider the general Box-Cox regression model in the form

$$y_t^{(\lambda_1)} = \beta_1 + \beta_2 x_{2t}^{(\lambda_2)} + \cdots + \beta_k x_{kt}^{(\lambda_k)} + \epsilon_t,\ \epsilon_t \sim iidN(0, \sigma^2),\ t = 1, \ldots, T \qquad (4.1)$$

where $y_t > 0$ and x_{it}, $i = 1, \ldots, k$, are assumed to be nonstochastic. Let the linear combination of the transformed explanatory variables be denoted by

$$\mu_t = \beta_1 + \beta_2 x_{2t}^{(\lambda_2)} + \cdots + \beta_k x_{kt}^{(\lambda_k)}$$

Then the Box-Cox model can be represented as

$$y_t^{(\lambda_1)} = \mu_t + \epsilon_t \tag{4.2}$$

The likelihood function of the usual Box-Cox model is

$$L(\cdot|\underline{y}, X) = (\frac{1}{\sqrt{2\pi\sigma^2}})^T \exp[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t^{(\lambda_1)} - \mu_t)^2] \prod_{t=1}^{T} y_t^{\lambda_1 - 1}$$

implying that the density of y_t is

$$g(y_t) = y_t^{\lambda_1 - 1}\frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t^{(\lambda_1)} - \mu_t)^2] \; (\lambda_1 \neq 0)$$

The integral of $g(y_t)$ over $\Re^+$ can be written

$$\int_0^\infty g(y_t)dy_t = \int_0^\infty y_t^{\lambda_1 - 1}\frac{1}{\sigma}\phi(\frac{y_t^{(\lambda_1)} - \mu_t}{\sigma})dy_t$$

where $\phi(\cdot)$ is the probability density function of a standard normal random variable. Let $z_t = (y_t^{(\lambda_1)} - \mu_t)/\sigma$. Then we obtain the following result:

If $\lambda_1 > 0$,

$$\begin{aligned}
\int_0^\infty g(y_t)dy_t &= \int_{-\frac{1}{\lambda_1}}^\infty \frac{1}{\sigma}\phi(z_t)dy_t^{(\lambda_1)} \\
&= \int_{-\frac{1+\lambda_1\mu_t}{\sigma\lambda_1}}^\infty \phi(z_t)dz_t \\
&= 1 - \Phi(-\frac{1+\lambda_1\mu_t}{\sigma\lambda_1}) \\
&= \Phi(\frac{1+\lambda_1\mu_t}{\sigma\lambda_1})
\end{aligned}$$

where $\Phi(\cdot)$ is the distribution function of a standard normal random variable.

If $\lambda_1 < 0$,

$$\begin{aligned}\int_0^{\infty} g(y_t)dy_t &= \int_{-\infty}^{-\frac{1}{\lambda_1}} \frac{1}{\sigma}\phi(z_t)dy_t^{(\lambda_1)} \\ &= \int_{-\infty}^{-\frac{1+\lambda_1\mu_t}{\sigma\lambda_1}} \phi(z_t)dz_t \\ &= \Phi(-\frac{1+\lambda_1\mu_t}{\sigma\lambda_1}) \\ &= 1-\Phi(\frac{1+\lambda_1\mu_t}{\sigma\lambda_1})\end{aligned}$$

Therefore, $g(y_t)$ is not a proper probability density function except when $\lambda_1 = 0$ under the usual Box-Cox transformation model.

Since the pre-transformed dependent variable y_t is truncated ($y_t > 0$), the transformed variable $y_t^{(\lambda_1)}$ is also truncated:

$$y_t^{(\lambda_1)} = y_t^* \quad \text{if } L < y_t^* < R$$

where y_t^* is an unobservable variable with distribution $N(\mu_t, \sigma^2)$; $R = -\frac{1}{\lambda_1}$ and $L = -\infty$ if $\lambda_1 < 0$; $R = +\infty$ and $L = -\frac{1}{\lambda_1}$ if $\lambda_1 > 0$. Therefore, the density of the transformed dependent variable is

$$\begin{aligned}&f(y_t^{(\lambda_1)}) \\ &\quad = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t^{(\lambda_1)}-\mu_t)^2][\Phi(R_t)-\Phi(L_t)]^{-1} \text{ if } L_t < y_t^{(\lambda_1)} < R_t \\ &\quad = 0 \quad \text{elsewhere}\end{aligned}$$

where $R_t = (R-\mu_t)/\sigma$ and $L_t = (L-\mu_t)/\sigma$. Let $\underline{\theta} = (\beta_1, \ldots, \beta_k, \lambda_1, \ldots, \lambda_k, \sigma^2)'$.

The log-likelihood function of the truncated Box-Cox model is written as

$$\begin{aligned}
\ell(\underline{\theta}; X, \underline{y}) &= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 \\
&\quad - \frac{1}{2\sigma^2}(\underline{y}^{(\lambda_1)} - X^{(\lambda_\cdot)}\underline{\beta})'(\underline{y}^{(\lambda_1)} - X^{(\lambda_\cdot)}\underline{\beta}) \\
&\quad - \sum_{t=1}^{T} \ln(\Phi(R_t) - \Phi(L_t)) + (\lambda_1 - 1)\sum_{t=1}^{T} \ln y_t \\
&= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\underline{\epsilon}'\underline{\epsilon} \\
&\quad - \sum_{t=1}^{T} \ln G_t + (\lambda_1 - 1)\sum_{t=1}^{T} \ln y_t \qquad (4.3)
\end{aligned}$$

where $\underline{y}^{(\lambda_1)}$ is a $T \times 1$ vector of the transformed dependent variable; $X^{(\lambda_\cdot)}$is a $T \times k$ matrix of independent variables with the Box-Cox transformation applied to each column vector except for the first, which is a constant vector; $G_t = \Phi(R_t) - \Phi(L_t)$; and $\underline{\epsilon} = \underline{y}^{(\lambda_1)} - X^{(\lambda_\cdot)}\underline{\beta}$. Consider the first derivatives for the log-likelihood function (4.3):

$$\begin{aligned}
\frac{\partial \ell}{\partial \underline{\beta}} &= \frac{1}{\sigma^2} X^{(\lambda_\cdot)'}\underline{\epsilon} - \sum \frac{1}{G_t}\frac{\partial G_t}{\partial \underline{\beta}} \\
\frac{\partial \ell}{\partial \lambda_1} &= -\frac{1}{\sigma^2}\underline{y}'_{\lambda_1}\underline{\epsilon} + \sum \ln y_t - \sum \frac{1}{G_t}\frac{\partial G_t}{\partial \lambda_1} \\
\frac{\partial \ell}{\partial \lambda_i} &= \frac{1}{\sigma^2}\beta_i \underline{x}'_{\lambda_i}\underline{\epsilon} - \sum \frac{1}{G_t}\frac{\partial G_t}{\partial \lambda_i}, \quad i > 1 \\
\frac{\partial \ell}{\partial \sigma^2} &= -\frac{T}{2}\sigma^2 + \frac{1}{2\sigma^4}\underline{\epsilon}'\underline{\epsilon} - \sum \frac{1}{G_t}\frac{\partial G_t}{\partial \sigma^2}
\end{aligned}$$

where

$$\underline{z}_{\lambda_i} = \frac{\partial \underline{z}_i^{(\lambda_i)}}{\partial \lambda_i} = \begin{cases} [(1 + \lambda_i \underline{z}_i^{(\lambda_i)}) \# \ln(1 + \lambda_i \underline{z}_i^{(\lambda_i)}) - \lambda_i \underline{z}_i^{(\lambda_i)}]/\lambda_i^2 & \text{if } \lambda_i \neq 0 \\ (\ln \underline{z}_i) \# (\ln \underline{z}_i)/2 & \text{if } \lambda_i = 0 \end{cases}$$

$$\underline{z}_i^{(\lambda_i)} = \begin{cases} \underline{y}^{(\lambda_1)} & \text{if } i = 1 \\ \underline{x}_i^{(\lambda_i)} & \text{if } i > 1 \end{cases}$$

$$
\begin{aligned}
\# &= \text{elementwise multiplication} \\
\frac{\partial G_t}{\partial \theta_i} &= \phi(R_t)\frac{\partial R_t}{\partial \theta_i} - \phi(L_t)\frac{\partial L_t}{\partial \theta_i} \\
\phi(\cdot) &= \text{probability density function of an } N(0,1) \text{ random variable.}
\end{aligned}
$$

Let

$$
\underline{z}_{\lambda_i\lambda_i} = \frac{\partial^2 \underline{z}^{(\lambda_i)}}{\partial \lambda_i^2}
$$

The second derivatives of the log-likelihood function are

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \underline{\beta}\partial \underline{\beta}'} &= -\frac{1}{\sigma^2}X^{(\lambda.)'}X^{(\lambda.)} - \sum \frac{1}{G_t}[\frac{\partial G_t}{\partial \underline{\beta}}\frac{\partial G_t}{\partial \underline{\beta}'} - G_t\frac{\partial^2 G_t}{\partial \underline{\beta}\partial \underline{\beta}'}] \\
\frac{\partial^2 \ell}{\partial \underline{\beta}\partial \lambda_1} &= -\frac{1}{\sigma^2}X^{(\lambda.)'}\underline{y}_{\lambda_1} - \sum \frac{1}{G_t}[\frac{\partial G_t}{\partial \underline{\beta}}\frac{\partial G_t}{\partial \lambda_1} - G_t\frac{\partial^2 G_t}{\partial \underline{\beta}\partial \lambda_1}] \\
\frac{\partial^2 \ell}{\partial \underline{\beta}\partial \lambda_i} &= -\frac{1}{\sigma^2}(\frac{\partial \beta_i}{\partial \underline{\beta}}\underline{x}'_{\lambda_i}\underline{\epsilon} - \beta_i X^{(\lambda.)'}\underline{x}_{\lambda_i}) - \sum \frac{1}{G_t}[\frac{\partial G_t}{\partial \underline{\beta}}\frac{\partial G_t}{\partial \lambda_i} - G_t\frac{\partial^2 G_t}{\partial \underline{\beta}\partial \lambda_i}],\ i = 2,\ldots,k \\
\frac{\partial^2 \ell}{\partial \underline{\beta}\partial \sigma^2} &= -\frac{1}{\sigma^4}X^{(\lambda.)'}\underline{\epsilon} - \sum \frac{1}{G_t}[\frac{\partial G_t}{\partial \underline{\beta}}\frac{\partial G_t}{\partial \sigma^2} - G_t\frac{\partial^2 G_t}{\partial \underline{\beta}\partial \sigma^2}] \\
\frac{\partial^2 \ell}{\partial \lambda_1^2} &= -\frac{1}{\sigma^2}(\underline{y}'_{\lambda_1}\underline{y}_{\lambda_1} + \underline{y}'_{\lambda_1\lambda_1}\underline{\epsilon}) - \sum \frac{1}{G_t}[(\frac{\partial G_t}{\partial \lambda_1})^2 - G_t\frac{\partial^2 G_t}{\partial \lambda_1^2}] \\
\frac{\partial^2 \ell}{\partial \lambda_1\partial \lambda_i} &= \frac{1}{\sigma^2}\beta_i\underline{y}'_{\lambda_1}\underline{x}_{\lambda_i} - \sum \frac{1}{G_t}[\frac{\partial G_t}{\partial \lambda_1}\frac{\partial G_t}{\partial \lambda_i} - G_t\frac{\partial^2 G_t}{\partial \lambda_1\partial \lambda_i}],\ i = 2,\ldots,k \\
\frac{\partial^2 \ell}{\partial \lambda_1\partial \sigma^2} &= \frac{1}{\sigma^4}\underline{y}'_{\lambda_1}\underline{\epsilon} - \sum \frac{1}{G_t}[\frac{\partial G_t}{\partial \lambda_1}\frac{\partial G_t}{\partial \sigma^2} - G_t\frac{\partial^2 G_t}{\partial \lambda_1\partial \sigma^2}] \\
\frac{\partial^2 \ell}{\partial \lambda_i^2} &= -\frac{1}{\sigma^2}(\beta_i^2\underline{x}'_{\lambda_i}\underline{x}_{\lambda_i} - \beta_i\underline{x}'_{\lambda_i\lambda_i}\underline{\epsilon}) - \sum \frac{1}{G_t}[(\frac{\partial G_t}{\partial \lambda_i})^2 - G_t\frac{\partial^2 G_t}{\partial \lambda_i^2}],\ i = 2,\ldots,k \\
\frac{\partial^2 \ell}{\partial \lambda_i\partial \lambda_j} &= -\frac{1}{\sigma^2}\beta_i\beta_j\underline{x}'_{\lambda_i}\underline{x}_{\lambda_j} - \sum \frac{1}{G_t}[\frac{\partial G_t}{\partial \lambda_i}\frac{\partial G_t}{\partial \lambda_j} - G_t\frac{\partial^2 G_t}{\partial \lambda_i\partial \lambda_j}],\ i \neq j \\
\frac{\partial^2 \ell}{\partial (\sigma^2)^2} &= -\frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\underline{\epsilon}'\underline{\epsilon} - \sum \frac{1}{G_t}[(\frac{\partial G_t}{\partial \sigma^2})^2 - G_t\frac{\partial^2 G_t}{\partial (\sigma^2)^2}]
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{\partial^2 G_t}{\partial \theta_i\partial \theta_j} &= \\
&\phi(R_t)[\frac{\partial^2 R_t}{\partial \theta_i\partial \theta_j} - R_t\frac{\partial R_t}{\partial \theta_i}\frac{\partial R_t}{\partial \theta_j}] - \phi(L_t)[\frac{\partial^2 L_t}{\partial \theta_i\partial \theta_j} - L_t\frac{\partial L_t}{\partial \theta_i}\frac{\partial L_t}{\partial \theta_j}],\ i,j = 1,\ldots,K
\end{aligned}
$$

Let $\underline{q}(\theta) = \frac{\partial \ell}{\partial \underline{\theta}}$, $Q(\underline{\theta}) = \frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'}$. The information matrix $-EQ(\underline{\theta}_0)$ is denoted by $\Im(\underline{\theta}_0)$. In addition, $Q_T(\underline{\theta}) = \frac{1}{T}Q(\underline{\theta})$ and $\Im_T(\underline{\theta}_0) = \frac{1}{T}\Im(\underline{\theta}_0)$. The dependence of the score vector and information matrix on the parameter vector will often be ignored, i.e., the score vector will be denoted by $\underline{q}$, rather than $\underline{q}(\underline{\theta}_0)$. When the score vector is evaluated at the ML estimator it will be denoted $\hat{\underline{q}} = \underline{q}(\hat{\underline{\theta}})$; when evaluated at the constrained ML vector it is $\tilde{\underline{q}} = \underline{q}(\tilde{\underline{\theta}})$.

4.3 Asymptotic Variabilities and Test Statistics

In this Section we discuss the variability of estimators and test procedures for linear restrictions. Three consistent estimators for the asymptotic covariance matrix are presented. The LM, LR and Wald statistics are shown to have equivalent χ^2-distribution under the null hypothesis. These three tests also have the same noncentral χ^2-distribution under local alternatives.

4.3.1 Asymptotically Equivalent Covariance Matrix Estimators

A Taylor series expansion of $\underline{q}$ about ML estimator $\hat{\underline{\theta}}$ yields

$$\underline{q} = \hat{\underline{q}} + Q(\underline{\theta}^*)(\underline{\theta}_0 - \hat{\underline{\theta}}) \tag{4.4}$$

where $\|\underline{\theta}^* - \underline{\theta}_0\| \leq \|\hat{\underline{\theta}} - \underline{\theta}_0\|$. Premultiplying both sides of (4.4) by $\frac{1}{\sqrt{T}}\Im_T^{-1/2}$, we obtain

$$\begin{aligned} \Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q} &= \Im_T^{-1/2}\frac{1}{\sqrt{T}}\hat{\underline{q}} + \Im_T^{-1/2}Q_T(\underline{\theta}^*)\sqrt{T}(\underline{\theta}_0 - \hat{\underline{\theta}}) \\ &= \Im_T^{1/2}\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) + \Im_T^{1/2}[\Im_T^{-1}Q_T(\underline{\theta}^*) + I]\sqrt{T}(\underline{\theta}_0 - \hat{\underline{\theta}}) \end{aligned}$$

since $\text{plim}\Im_T^{-1}Q_T(\underline{\theta}^*) = -I$ and $\Im_T^{-1}Q_T(\underline{\theta}^*) + I = o_p(1)$. According to the Central Limit Theorem,

$$\Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q} \xrightarrow{d} \underline{z}, \quad \underline{z} \sim N(\underline{0}, I_K)$$

Theorem 4.1 *Let $\{\underline{x}_n\}$ be a sequence of random finite dimensional vectors. If $\underline{x}_n \xrightarrow{d} \underline{x}$ for some random vector, then $\underline{x}_n = O_p(1)$.*

Proof: [White (1984) p. 63]

Therefore, $\Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q} = O_p(1)$. Equation (4.4) can be rewritten as

$$\Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q} = \Im_T^{1/2}\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) + o_p(1) \tag{4.5}$$

which implies that

$$\Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q} \overset{A}{=} \Im_T^{1/2}\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0)$$

since the second term of the right hand side of (4.5) is $o_p(1)$ and thus negligible relative to $\Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q}$ and $\Im_T^{1/2}\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0)$, which are $O_p(1)$. Therefore,

$$\Im_T^{1/2}\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) \overset{a}{\sim} N(\underline{0}, I_K) \tag{4.6}$$

The ML estimator of the covariance matrix is $V_1 = \Im(\hat{\underline{\theta}})^{-1}$, since the true asymptotic covariance of $\hat{\underline{\theta}}$ is $\Im^{-1}$. The other two consistent estimators for the asymptotic variance of $\hat{\underline{\theta}}$ are the negative of the Hessian, $V_2 = -Q(\hat{\underline{\theta}})^{-1}$, and the outer product form $V_3 = [\sum_{t=1}^{T}\frac{\partial \ell_t}{\partial \underline{\theta}}\frac{\partial \ell_t}{\partial \underline{\theta}'}|_{\hat{\underline{\theta}}}]^{-1}$ (Berndt, Hall, Hall and Hausman (BHHH)).

Theorem 4.2 (Chebyshev)

Suppose $X_1, \ldots, X_T$ *are independent random variables such that* $E(X_t) = \mu_t$ *and* $V(X_t) = \sigma_t^2 < \infty$. *Then for any* $\varepsilon > 0$,

$$\lim_{T\to\infty} P(|\frac{1}{T}\sum X_t - \frac{1}{T}\sum \mu_t| < \varepsilon) = 1$$

The log-likelihood function (4.3) can be written as

$$\ell(\underline{\theta}) = \sum_{t=1}^{T} \ln\ f_t(\underline{\theta}) = \sum_{t=1}^{T} \ell_t(\underline{\theta})$$

where the $\ell_t(\underline{\theta})$ are independent. Using Chebyshev's theorem, it can be shown that

$$\text{plim}\frac{1}{T}\sum \frac{\partial \ell_t}{\partial \underline{\theta}} \frac{\partial \ell_t}{\partial \underline{\theta}'}|_{\hat{\underline{\theta}}} = \lim_{T\to\infty} E\frac{1}{T}\frac{\partial \ell}{\partial \underline{\theta}} \frac{\partial \ell}{\partial \underline{\theta}'}|_{\underline{\theta}_0}$$

The density of y_t is regular, so that

$$\Im = E\frac{\partial \ell}{\partial \underline{\theta}} \frac{\partial \ell}{\partial \underline{\theta}'}|_{\underline{\theta}_0}$$

Therefore, the outer product of variance estimator V_3 is a consistent estimator of the asymptotic covariance matrix of the ML estimator $\hat{\underline{\theta}}$. Similarly, using Chebyshev's theorem, we have

$$\text{plim}[-Q_T(\hat{\underline{\theta}})]^{-1}\Im_T = I$$

Thus V_2 is a consistent estimator for $\Im^{-1}$.

Griffiths, Hill and Pope (1987) investigated the small sample properties of these three asymptotically equivalent covariance estimators within the context of the probit regression model. The average variance estimates based on V_1 and

V_2 were very similar in magnititude, while the average variance estimate based on the outer product form (BHHH) V_3 was larger than that based on V_1 or V_2. The variance estimate V_3, however, seems to be a better approximation to the finite sample variability. A similar study in the context of simultaneous equations was performed by Calzolari and Panattoni (1988). Their Monte Carlo experiment showed that the average standard errors computed from the Hessian matrix are smaller than those (in all but one case) computed from the outer product form.

The question we are concerned with is how accurately the asymptotic variance estimators approximate the small sample variabilities within the framework of the truncated Box-Cox model. The finite sample properties of the consistent estimators for the asymptotic ML covariance matrix estimators (V_2 and V_3) will also be investigated.

4.3.2 Asymptotically Equivalent Test Statistics

We assume that the null hypothesis to be tested takes the linear form:

$$H_0 : R\underline{\theta} - \underline{r} = \underline{0}$$

where R is a $J \times K$ matrix of rank $J(J < K)$. The constrained ML estimation is based on the maximand with Lagrange multiplier vector $\underline{\mu}$:

$$M(\underline{\theta}, \underline{\mu}) = \ell(\underline{\theta}) - \underline{\mu}'(R\underline{\theta} - \underline{r})$$

where $\underline{\mu} \in R^J$. The constrained ML estimator $\tilde{\underline{\theta}}$ is the solution of the Lagrange first-order necessary conditions:

$$\tilde{\underline{q}} - R'\tilde{\underline{\mu}} = \underline{0} \tag{4.7}$$

$$R\tilde{\underline{\theta}} - \underline{r} = \underline{0} \tag{4.8}$$

We consider a Taylor series expansion of (4.7) about $\underline{\theta}_0$:

$$\underline{q} + Q(\underline{\theta}^*)(\tilde{\underline{\theta}} - \underline{\theta}_0) - R'\tilde{\underline{\mu}} = \underline{0} \tag{4.9}$$

When we assume that H_0 is true, (4.8) is

$$R(\tilde{\underline{\theta}} - \underline{\theta}_0) = \underline{0} \tag{4.10}$$

Premultiplying (4.9) by $\frac{1}{\sqrt{T}}(R\Im_T^{-1}R')^{-1}R\Im_T^{-1}$,

$$(R\Im_T^{-1}R')^{-1}R\Im_T^{-1}\frac{1}{\sqrt{T}}\underline{q} + (R\Im_T^{-1}R')^{-1}R\Im_T^{-1}Q_T(\underline{\theta}^*)\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) - \frac{1}{\sqrt{T}}\tilde{\underline{\mu}} = \underline{0}$$

Since $\text{plim}\Im^{-1}Q(\underline{\theta}^*) = -I$ and $R(\tilde{\underline{\theta}} - \underline{\theta}_0) = \underline{0}$,

$$(R\Im_T^{-1}R')^{-1}R\Im_T^{-1}Q_T(\underline{\theta}^*)\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) \stackrel{A}{=} -(R\Im_T^{-1}R')^{-1}\sqrt{T}R(\tilde{\underline{\theta}} - \underline{\theta}_0) = \underline{0}$$

Consequently,

$$\frac{1}{\sqrt{T}}\tilde{\underline{\mu}} \stackrel{A}{=} (R\Im_T^{-1}R')^{-1}R\Im_T^{-1}\frac{1}{\sqrt{T}}\underline{q}$$

Therefore, we have the result that

$$(R\Im_T^{-1}R')^{1/2}\frac{1}{\sqrt{T}}\tilde{\underline{\mu}} \stackrel{a}{\sim} N(\underline{0}, I_J) \tag{4.11}$$

if H_0 is true.

The Lagrange multiplier statistic, which is in a form of Rao's score test statistic, is given by

$$LM = \tilde{\underline{q}}' \tilde{\Im}^{-1} \tilde{\underline{q}}$$

From (4.7), we know that $\tilde{\underline{q}} = R'\tilde{\underline{\mu}}$. Thus the LM test can be represented in the form that was suggested by Aitchison and Silvey (1959):

$$LM = \tilde{\underline{\mu}}' R \tilde{\Im}^{-1} R' \tilde{\underline{\mu}}$$

Using (4.11) and the relationship that $\text{plim} \Im_T(\tilde{\underline{\theta}})^{-1} \Im_T(\underline{\theta}_0) = I$, it follows that

$$LM = [(R\Im_T^{-1}R')^{1/2} \frac{1}{\sqrt{T}} \tilde{\underline{\mu}}]' [(R\Im_T^{-1}R')^{1/2} \frac{1}{\sqrt{T}} \tilde{\underline{\mu}}] \overset{a}{\sim} \chi^2_{(J)}$$

if H_0 is true.

The likelihood ratio test statistic is obtained by comparing the maximum log-likelihood values evaluated at the constrained and unconstrained ML estimates:

$$LR = -2(\ell(\tilde{\underline{\theta}}) - \ell(\hat{\underline{\theta}}))$$

The Taylor series expansion of $\ell(\tilde{\underline{\theta}})$ about $\hat{\underline{\theta}}$ will lead to the following relationship:

$$\ell(\tilde{\underline{\theta}}) = \ell(\hat{\underline{\theta}}) + \hat{\underline{q}}(\tilde{\underline{\theta}} - \hat{\underline{\theta}}) + \frac{1}{2}(\tilde{\underline{\theta}} - \hat{\underline{\theta}})' Q(\hat{\underline{\theta}})(\tilde{\underline{\theta}} - \hat{\underline{\theta}}) + O_p(T^{-1/2}) \qquad (4.12)$$

Since both of $\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0)$ and $\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0)$ are asymptotically normally distributed, they are $O_p(1)$ from Theorem 4.1. Then $(\tilde{\underline{\theta}} - \hat{\underline{\theta}}) = (\tilde{\underline{\theta}} - \underline{\theta}_0) - (\hat{\underline{\theta}} - \underline{\theta}_0)$ is at most $O_p(T^{-1/2})$ under the null hypothesis. Therefore, (4.12) is represented by

$$\ell(\tilde{\underline{\theta}}) - \ell(\hat{\underline{\theta}}) \overset{A}{=} \frac{1}{2}(\tilde{\underline{\theta}} - \hat{\underline{\theta}})' Q(\hat{\underline{\theta}})(\tilde{\underline{\theta}} - \hat{\underline{\theta}})$$

since $\hat{\underline{q}} = \underline{0}$ and the term of $O_p(T^{-1/2})$ is negligible compared to that of $O_p(1)$. Then LR is asymptotically equivalent to

$$T(\tilde{\underline{\theta}} - \hat{\underline{\theta}})' Q_T(\hat{\underline{\theta}})(\tilde{\underline{\theta}} - \hat{\underline{\theta}}) \tag{4.13}$$

The vector $\tilde{\underline{q}}$ can be linearized around $\hat{\underline{\theta}}$:

$$\tilde{\underline{q}} = \hat{\underline{q}} + Q(\hat{\underline{\theta}})(\tilde{\underline{\theta}} - \hat{\underline{\theta}}) + O_p(1)$$

Thus,

$$\frac{1}{\sqrt{T}}\tilde{\underline{q}} \stackrel{A}{=} -\Im_T \sqrt{T}(\tilde{\underline{\theta}} - \hat{\underline{\theta}}) \tag{4.14}$$

Combining (4.7) and (4.14), we have

$$\frac{1}{\sqrt{T}} R' \tilde{\underline{\mu}} = -\Im_T \sqrt{T}(\tilde{\underline{\theta}} - \hat{\underline{\theta}})$$

Therefore, the LR test (4.13) can be written as

$$\begin{aligned} LR &\stackrel{A}{=} \frac{1}{T}\tilde{\underline{\mu}}' R \Im_T^{-1} R' \tilde{\underline{\mu}} \\ &= [(R\Im_T^{-1}R')^{1/2}\frac{1}{\sqrt{T}}\tilde{\underline{\mu}}]'[(R\Im_T^{-1}R')^{1/2}\frac{1}{\sqrt{T}}\tilde{\underline{\mu}}] \stackrel{a}{\sim} \chi^2_{(J)} \end{aligned}$$

if H_0 is true.

The Wald statistic is based on the idea that $R\hat{\underline{\theta}} - \underline{r} = \underline{0}$ if H_0 is true. The Wald test is given by

$$W = (R\hat{\underline{\theta}} - \underline{r})'[R\hat{\Im}^{-1}R']^{-1}(R\hat{\underline{\theta}} - \underline{r})$$

When H_0 is true, $R\hat{\underline{\theta}} - \underline{r} = R(\hat{\underline{\theta}} - \underline{\theta}_0)$. Then

$$\begin{aligned} W &= T(\hat{\underline{\theta}} - \underline{\theta}_0)' R'[R\Im_T^{-1}R']^{-1} R(\hat{\underline{\theta}} - \underline{\theta}_0) \\ &\stackrel{A}{=} [(R\Im_T^{-1}R')^{-1/2} R\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0)]'[(R\Im_T^{-1}R')^{-1/2} R\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0)] \stackrel{a}{\sim} \chi^2_{(J)} \end{aligned}$$

under the null hypothesis.

The LM, LR and Wald tests have an asymptotic noncentral χ^2-distribution under a sequence of local alternatives $H_T : R\underline{\theta} - \underline{r} = \underline{\delta}/\sqrt{T}$ such that $\text{plim}\sqrt{T}(R\underline{\theta} - \underline{r}) = \underline{\delta}$ for some vector $\underline{\delta}$. The noncentrality parameter can be derived only by consideration of H_T such that

$$\lambda = \underline{\delta}'(R\Im_T^{-1}R')^{-1}\underline{\delta}$$

The arbitrary constant vector $\underline{\delta}$ needs to be small enough that the asymptotic power is less than one. Suppose $R = I_K$ and $\underline{r} = \underline{\theta}_1$. The asymptotic power function for any fixed level α is denoted by

$$\mathcal{P}(\underline{\theta}_1) = P[\chi^2_{(J,\lambda)} > \chi^2_{(J)}(\alpha);\ \underline{\theta}_1]$$

where $\underline{\theta}_1 \neq \underline{\theta}_0$. For a fixed $\underline{\theta}_1$, the asymptotic power functions for LM, LR and Wald tests satisfy

$$\lim_{T\to\infty} \mathcal{P}(\underline{\theta}_1) = 1$$

We will consider an example of a noncentral Wald test. The local alternatives lead to the relationship:

$$\sqrt{T}(R\hat{\underline{\theta}} - \underline{r}) = \sqrt{T}R(\hat{\underline{\theta}} - \underline{\theta}_0) + \underline{\delta}$$

Since $\underline{\delta}$ is a constant vector,

$$\sqrt{T}(R\hat{\underline{\theta}} - \underline{r}) \overset{a}{\sim} N(\underline{\delta}, \lim_{T\to\infty} R\Im_T^{-1}R')$$

The Wald test under H_T is represented by

$$W_1 = (R\underline{\hat{\theta}} - \underline{r})'[R\hat{\Im}^{-1}R']^{-1}(R\underline{\hat{\theta}} - \underline{r}) \overset{a}{\sim} \chi^2_{(J,\lambda)}$$

where $\lambda = \underline{\delta}'(R\Im_T^{-1}R')^{-1}\underline{\delta}$. Suppose $R\underline{\theta} \neq \underline{r}$. Then the Wald statistic W_1 is $O_p(T)$ since $R\underline{\hat{\theta}} - \underline{r} = O_p(1)$ and $(R\hat{\Im}^{-1}R')^{-1} = O_p(T)$. It is expected that W_1 will exceed any constant critical value with probability one when T tends to infinity, i.e.,

$$\lim_{T\to\infty} P[W_1 > \chi^2_{(J)}(\alpha)] = 1$$

if $R\underline{\theta} \neq \underline{r}$ [Godfrey (1988), p 17].

4.4 A Monte Carlo Simulation

We have discussed asymptotically equivalent covariance matrix estimators and test statistics. In this Section we explain the model and simulation design concerning test procedures in the context of the general Box-Cox transformation model. First, standard errors computed from the Hessian matrix and outer products of the first derivatives are compared to the root mean square error (RMSE), which represents the finite sample variability. The information matrix is obtained using numerical integration (GAUSS function INTQUAD1). The ratio of the standard error computed from the information matrix over the RMSE is a measure of the closeness of the small sample approximation to the asymptotic variance. Second, the sampling distributions of the LM, LR and Wald statistics for testing functional forms of the model are examined using the Kolmogorov-Smironov (KS) test when

the null hypothesis is true. The real (empirical) size of the asymptotic χ^2-statistics is computed as the proportion of trials that the given statistic exceeds the critical values corresponding to nominal sizes ($\alpha = 0.1$, 0.5 and 1.0). Finally, we are concerned with the power function of the asymptotic χ^2-statistics. The estimated power is used to see how well the asymptotic power function performs in small samples.

The model is specified as follows:

$$y_t^{(\lambda_1)} = sign(\lambda_1)10.0 + sign(\lambda_1)1.5x_{2t}^{(0.1)} - sign(\lambda_1)0.5x_{3t}^{(1.0)} + \epsilon_t, \quad \epsilon \sim N(0, \sigma^2)$$

where $\lambda_1 = (-2.0, -1.0, -0.5, 0.1, 0.5, 1.0, 2.0)$ and $\sigma^2 = (0.1, 0.5)$. We can only observe the truncated value of $y_t^{(\lambda_1)}$ since $y_t > 0$. The error disturbances ϵ_t are generated by the GAUSS random number generator RNDNS such that $y_t^{(\lambda_1)} > -\frac{1}{\lambda_1}$ if $\lambda_1 > 0$; $y_t^{(\lambda_1)} < -\frac{1}{\lambda_1}$ if $\lambda_1 < 0$. The model is chosen to have two explanatory variables whose values are generated from linear combinations of uniform random numbers (GAUSS function RNDUS) such that $corr(x_{2t}, x_{3t}) = 0.2$:

$$\begin{aligned} x_{2t} &= 12 + 4u_t \\ x_{3t} &= 15 + u_t + 2\sqrt{6}v_t \end{aligned}$$

where u_t and v_t are uniformly distributed on the interval $U(-\sqrt{3}, \sqrt{3})$. The selected variances of the error term ($\sigma^2 = 0.1$ and 0.5) account for approximately 2 % and 6 % of the variation of the right-hand side of the Box-Cox model, respectively.

Each Box-Cox model is estimated 1000 times for samples of size $T = 30$ and $T = 60$. For the analysis of power properties, the log-linear model is tested, i.e.,

$H_0 : \lambda_1 = \lambda_2 = 0$ and $\lambda_3 = 1.0$. Let $\bar{\theta}_{ij}$ be the ith parameter estimator in the jth replication of simulation. The bias and RMSE are defined by

$$\begin{aligned} \text{bias} &= \frac{\sum_{j=1}^{N} \bar{\theta}_{ij}}{N} - \theta_i \\ \text{RMSE} &= \sqrt{\sum_{j=1}^{N} (\bar{\theta}_{ij} - \theta_i)^2 / N} \end{aligned}$$

where θ_i is the ith parameter and N represents the number of simulation replicates. Let $\underline{\theta}_1 = (\beta_1, \beta_2, \beta_3, \sigma^2)'$ and $\underline{\theta}_2 = (\lambda_1, \lambda_2, \lambda_3)'$. The test statistics are computed under $H_0 : \underline{\theta}_2 = \underline{\theta}_2^0$:

$$\begin{aligned} \text{Wald}_1 &= (\hat{\underline{\theta}}_2 - \underline{\theta}_2^0)'[-Q_{22}(\hat{\underline{\theta}}) + Q_{21}(\hat{\underline{\theta}}) Q_{22}^{-1}(\hat{\underline{\theta}}) Q_{12}(\hat{\underline{\theta}})](\hat{\underline{\theta}}_2 - \underline{\theta}_2^0) \\ \text{Wald}_2 &= (\hat{\underline{\theta}}_2 - \underline{\theta}_2^0)'[P_{22}(\hat{\underline{\theta}}) - P_{21}(\hat{\underline{\theta}}) P_{22}^{-1}(\hat{\underline{\theta}}) P_{12}(\hat{\underline{\theta}})](\hat{\underline{\theta}}_2 - \underline{\theta}_2^0) \\ \text{LM} &= \underline{q}(\tilde{\underline{\theta}})' P(\tilde{\underline{\theta}}) \underline{q}(\tilde{\underline{\theta}}) \\ \text{LR} &= -2(\ell(\tilde{\underline{\theta}}) - \ell(\hat{\underline{\theta}})) \end{aligned}$$

where

$$\begin{aligned} Q_{ij}(\hat{\underline{\theta}}) &= \frac{\partial^2 \ell}{\partial \underline{\theta}_i \partial \underline{\theta}_j'} |_{\hat{\underline{\theta}}}, \quad i, j = 1, 2 \\ P_{ij}(\hat{\underline{\theta}}) &= \sum_{t=1}^{T} \frac{\partial \ell_t}{\partial \underline{\theta}_i} \frac{\partial \ell_t}{\partial \underline{\theta}_j'} |_{\hat{\underline{\theta}}}, \quad i, j = 1, 2 \\ P &= \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \end{aligned}$$

4.5 Results

4.5.1 Bias

The Monte Carlo results on the biases of the ML estimators are given in Table 4.1. In general, the magnitudes of the biases of the linear parameters are greater

than those of the power parameters. The bias of the power parameter estimator is negligible. When the sample size increases, the magnitudes of the biases of all estimators which are greater than unity are reduced at $T = 30$.

The estimator $\hat{\beta}_2$ is substantially biased when $T = 30$ and $\sigma^2 = 0.5$. The bias of $\hat{\beta}_2$ is the largest among parameter estimators for $T = 30$ and $T = 60$. The linear parameter estimates are greatly biased when $T = 30$ and $\sigma^2 = 0.5$. $\hat{\beta}_2$ is biased downward when $\lambda_1 < 0$ and biased upward when $\lambda_1 > 0$. The reverse is true for $\hat{\beta}_1$. Overall, the direction of bias is unchanged, with the magnitude of bias greater than unity as the sample size increases. Concerning λ_1, the bias is larger at $\lambda_1 = \pm 2$ than at $|\lambda_1| < 2.0$ when $\sigma^2 = 0.1$.

4.5.2 RMSE and Standard Errors

The parameter RMSE's and standard errors are reported in Tables 4.2–4.5. The standard errors computed from the information matrix greatly understate the RMSE's except for $\hat{\lambda}_2$. The standard errors of the consistent covariance matrix estimators (negative of the inverted Hessian and inverted outer products) seem to be a good approximation to the RMSE in the case of power parameter estimation. The standard errors from the outer products are greater than those from the Hessian matrix.

The larger the variance of the error disturbances, the larger are the finite sample variabilities. RMSE's are smaller at $T = 60$ than at $T = 30$. Linear parameter estimators have large RMSE's when $T = 30$ and $\sigma^2 = 0.5$. On the whole, the

standard errors computed from the outer products gives a better approximation to the finite sample variability than those from the Hessian matrix.

4.5.3 Empirical Distribution and Real Size of the Tests

The one-sample KS test is concerned with the quality of approximation of a sampling distribution to a theoretical distribution which would be expected under the null hypothesis. Critical values for one-tailed tests at level α are approximated by

$$C_\alpha \simeq [\ln(2/\alpha)/2N]^{1/2}$$

where C_α satisfies $P(KS > C_\alpha) = \alpha$ [Morimune (1989)]. When $\alpha = 0.05$ and $N = 1000$, $C_\alpha = 0.043$. Results of the KS test for the test statistics Wald_1, Wald_2, LM and LR are reported in Tables 4.6 and 4.7.

Obviously, LR tests give the best approximation to the central χ^2_3-distribution when H_0 is true. The KS test statistics of the Wald_1, Wald_2 and LR are reduced in magnitude as the sample size increases. When $\sigma^2 = 0.1$, the Wald statistic using the Hessian matrix (Wald_1) generally gives a better approximation to the χ^2_3-distribution than the Wald statistic using outer products (Wald_2). However, the quality of approximation of Wald_1 and Wald_2 to the χ^2_3-distribution is virtually indistinguishable when $\sigma^2 = 0.5$. The outer product variant of the LM test statistic is shown to be a poor approximation to the χ^2_3-distribution in small samples ($T = 30$ and $T = 60$).

Real sizes of the tests are given in Tables 4.8–4.11. In general, the real size

of the LR test is similar to the nominal size. When the variance of the error disturbances is 0.1, Wald_1 slightly performs better than Wald_2 and the real sizes of Wald_1 look similar to those of LR. The small sample LM test shows a large rejection rate under the null hypothesis.

4.5.4 Power Properties

It is assumed that we are to test $H_0 : \underline{\theta}_2 = (0,0,1)'$ against local alternatives $H_T : \underline{\theta}_2 = (0,0,1)' + \underline{\delta}/\sqrt{T}$. The constant vector $\underline{\delta}$ is chosen to be $\sqrt{T}(0.1, 0.1, 0.0)'$ since we estimated the true model with $\lambda_1 = \lambda_2 = 0.1$ and $\lambda_3 = 1.0$. The noncentrality parameter is

$$\lambda = \underline{\delta}' \frac{1}{T} (\Im_{22} - \Im_{21}\Im_{11}^{-1}\Im_{12})\underline{\delta}$$

where $\Im_{22}$ is the 3×3 submatrix of the information matrix corresponding to the ML estimator $\hat{\underline{\theta}}_2$ and $\Im_{11}$ is the 4×4 submatrix related to the other ML estimators. We obtained the noncentrality parameters as follows:

$$\lambda = 24.390 \quad \text{if } T = 30 \text{ and } \sigma^2 = 0.1$$

$$\lambda = 4.923 \quad \text{if } T = 30 \text{ and } \sigma^2 = 0.5$$

$$\lambda = 50.260 \quad \text{if } T = 60 \text{ and } \sigma^2 = 0.1$$

$$\lambda = 9.794 \quad \text{if } T = 60 \text{ and } \sigma^2 = 0.5$$

The asymptotic power function at level α is $P[\chi^2_{(3,\lambda)} > C_\alpha]$ where C_α is the α-level critical value of the central χ^2_3-distribution.

Empirical results are reported in Table 4.12. For $T = 30$, the asymptotic powers exceed the estimated powers. On the other hand, the estimated powers are greater than the asymptotic powers when $T = 60$ and $\sigma^2 = 0.1$. The estimated power increases with the increasing sample size, as expected. The LM statistic rejects H_0 most frequently. Wald_1 is more powerful than Wald_2 and LR in all cases and LR performs better than Wald_2. In our limited experiments, the asymptotic power function does not provide a good estimate of the small sample power.

4.6 Conclusions

The observed results lead us to conclude that:

1. In general, the bias of the power parameter estimator is negligible. The bias of the estimator of the parameter that is related to the variable of the nonlinear transformation ($\lambda_1 = 0.1$) is substantial in small samples. When the variance of error disturbances is small ($\sigma^2 = 0.1$), the magnitude of the bias is larger for $|\lambda_1| = 2.0$ than for $|\lambda_1| < 2.0$.

2. The finite sample variabilities (RMSE's) of the parameter estimators are larger for $\sigma^2 = 0.5$ than for $\sigma^2 = 0.1$. The RMSE's are reduced as the sample size increases.

3. The asymptotic standard error from the information matrix is not a good measure of finite sample variability. The degree of agreement between the

RMSE and the asymptotic standard error increases with increasing sample size.

4. The standard error computed from outer products of the first derivatives is a better approximation to the RMSE than the standard error obtained from the Hessian matrix.

5. According to the results of the Kolmogorov-Smironov tests, the LR statistic provides a better approximation to the asymptotic χ^2_3-distribution than either of the other two. The LM statistic based on outer products is indicative of the poor approximation to the asymptotic distribution. The estimated size is closer to the corresponding nominal size for the LR statistic than for any other statistics. The small sample LM test statistics are not reliable under the null hypothesis, based on the Kolmogorov-Smironov and empirical size results.

6. The asymptotic power function is not a good approximation to the small sample power for either $T = 30$ or $T = 60$. The estimated power increases as the sample size increases. The LM test turns out to be the most powerful test based on our experiments.

Table 4.1: Bias of MLE

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
$\lambda_1 =$ **-2.0**							
σ^2 =0.1	T=30	-1.220	-2.198	0.453	-0.047	-0.017	-0.030
	T=60	-0.811	-0.894	0.216	-0.011	0.007	-0.006
σ^2 =0.5	T=30	3.434	-20.349	2.003	-0.008	0.066	0.013
	T=60	0.018	-5.377	0.631	0.023	0.064	0.015
$\lambda_1 =$ **-1.0**							
σ^2 =0.1	T=30	-0.374	-1.490	0.316	0.004	0.013	0.006
	T=60	-0.436	-0.536	0.145	0.006	0.033	0.001
σ^2 =0.5	T=30	8.009	-31.480	2.089	0.003	0.060	0.006
	T=60	1.683	-6.425	0.465	0.028	0.056	0.014
$\lambda_1 =$ **-0.5**							
σ^2 =0.1	T=30	0.129	-2.001	0.331	0.005	0.003	0.015
	T=60	-0.010	-0.647	0.122	0.007	0.005	0.008
σ^2 =0.5	T=30	8.845	-29.008	1.725	0.019	0.064	0.033
	T=60	1.733	-5.566	0.429	0.011	0.020	0.006
$\lambda_1 =$ **0.1**							
σ^2 =0.1	T=30	-0.303	1.815	-0.273	0.000	-0.008	0.003
	T=60	-0.049	0.692	-0.109	0.001	-0.000	-0.003
σ^2 =0.5	T=30	-6.997	22.511	-1.499	-0.004	0.025	0.021
	T=60	-3.316	7.936	-0.381	-0.005	0.013	0.008
$\lambda_1 =$ **0.5**							
σ^2 =0.1	T=30	0.186	1.499	-0.284	-0.002	0.029	0.018
	T=60	-0.223	0.568	-0.078	-0.014	0.005	0.021
σ^2 =0.5	T=30	-8.387	27.965	-1.564	-0.022	0.013	0.058
	T=60	-1.824	5.565	-0.433	-0.015	0.030	0.015
$\lambda_1 =$ **1.0**							
σ^2 =0.1	T=30	0.396	1.821	-0.348	0.000	0.017	-0.002
	T=60	-0.032	0.815	-0.130	-0.013	-0.041	0.009
σ^2 =0.5	T=30	-5.447	28.577	-3.024	0.003	0.065	-0.007
	T=60	-1.076	6.311	-0.592	-0.017	0.074	0.019
$\lambda_1 =$ **2.0**							
σ^2 =0.1	T=30	1.386	2.431	-0.503	0.061	-0.013	-0.019
	T=60	1.096	0.736	-0.224	0.027	0.012	-0.009
σ^2 =0.5	T=30	-5.533	26.012	-2.252	-0.013	0.082	0.004
	T=60	0.008	6.133	-0.744	0.002	0.095	0.000

Table 4.2: RMSE and Standard Error of MLE (T=30 σ^2=0.1)

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
λ_1 = **-2.0**						
RMSE	8.763	7.208	1.319	0.370	0.519	0.222
SE_1/RMSE	0.303	0.249	0.186	0.221	0.983	0.759
SE_2/RMSE	0.863	0.628	0.705	1.082	0.988	0.976
SE_3/RMSE	1.121	0.751	0.937	1.368	1.064	1.223
λ_1 = **-1.0**						
RMSE	6.848	4.857	1.180	0.209	0.512	0.238
SE_1/RMSE	0.390	0.370	0.214	0.263	0.996	0.716
SE_2/RMSE	0.906	0.772	0.687	1.144	0.953	1.050
SE_3/RMSE	1.028	0.804	0.794	1.246	0.998	1.123
λ_1 = **-0.5**						
RMSE	7.242	7.325	1.316	0.134	0.541	0.262
SE_1/RMSE	0.372	0.245	0.198	0.294	0.944	0.663
SE_2/RMSE	0.804	0.555	0.578	0.955	0.899	0.898
SE_3/RMSE	0.917	0.628	0.677	1.090	0.925	1.001
λ_1 = **0.1**						
RMSE	5.447	6.261	0.897	0.056	0.523	0.263
SE_1/RMSE	0.502	0.287	0.316	0.412	0.977	0.705
SE_2/RMSE	0.864	0.581	0.702	0.878	0.928	0.884
SE_3/RMSE	1.010	0.691	0.776	0.965	1.029	0.982
λ_1 = **0.5**						
RMSE	6.507	6.097	1.107	0.131	0.524	0.252
SE_1/RMSE	0.414	0.295	0.235	0.300	0.974	0.691
SE_2/RMSE	0.837	0.547	0.643	0.969	0.919	0.919
SE_3/RMSE	1.038	0.647	0.745	1.189	1.033	1.056
λ_1 = **1.0**						
RMSE	7.616	6.311	1.294	0.211	0.527	0.235
SE_1/RMSE	0.351	0.285	0.195	0.261	0.968	0.724
SE_2/RMSE	0.851	0.605	0.621	1.094	0.913	1.006
SE_3/RMSE	0.927	0.662	0.740	1.109	0.957	1.100
λ_1 = **2.0**						
RMSE	9.232	8.212	1.516	0.376	0.521	0.229
SE_1/RMSE	0.287	0.219	0.162	0.217	0.980	0.735
SE_2/RMSE	0.829	0.571	0.634	1.057	0.978	0.942
SE_3/RMSE	1.031	0.675	0.770	1.198	1.048	1.105

SE_1 = standard error from the information matrix
SE_2 = standard error from the Hessian matrix
SE_3 = standard error from the outer product matrix

Table 4.3: RMSE and Standard Error of MLE (T=30 σ^2=0.5)

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
λ_1 = **-2.0**						
RMSE	44.738	111.267	9.342	0.560	1.135	0.454
SE_1/RMSE	0.132	0.035	0.063	0.357	0.967	0.886
SE_2/RMSE	0.510	0.322	0.402	1.141	0.945	0.830
SE_3/RMSE	1.323	1.410	0.563	1.400	1.052	1.148
λ_1 = **-1.0**						
RMSE	62.085	156.223	8.782	0.358	1.166	0.488
SE_1/RMSE	0.098	0.026	0.065	0.374	0.979	0.785
SE_2/RMSE	0.445	0.320	0.357	0.996	0.905	0.808
SE_3/RMSE	0.826	0.645	0.571	1.153	1.021	1.004
λ_1 = **-0.5**						
RMSE	53.282	138.574	9.858	0.205	1.206	0.495
SE_1/RMSE	0.115	0.029	0.060	0.457	0.946	0.791
SE_2/RMSE	0.580	0.490	0.352	1.044	0.890	0.899
SE_3/RMSE	0.786	0.632	0.391	1.188	0.927	0.975
λ_1 = **0.1**						
RMSE	39.349	101.126	6.865	0.090	1.187	0.526
SE_1/RMSE	0.155	0.040	0.092	0.568	0.962	0.787
SE_2/RMSE	0.650	0.485	0.417	1.008	0.925	0.978
SE_3/RMSE	0.826	0.652	0.456	1.100	0.985	0.994
λ_1 = **0.5**						
RMSE	50.461	130.896	10.274	0.212	1.189	0.504
SE_1/RMSE	0.121	0.031	0.058	0.442	0.960	0.778
SE_2/RMSE	0.457	0.325	0.232	0.977	0.889	0.858
SE_3/RMSE	0.916	0.777	0.337	1.173	1.026	0.990
λ_1 = **1.0**						
RMSE	85.687	229.848	17.994	0.350	1.165	0.500
SE_1/RMSE	0.071	0.017	0.032	0.382	0.979	0.766
SE_2/RMSE	0.310	0.203	0.209	1.024	0.919	0.798
SE_3/RMSE	0.559	0.396	0.312	1.139	1.132	1.073
λ_1 = **2.0**						
RMSE	45.191	120.574	15.121	0.575	1.179	0.447
SE_1/RMSE	0.131	0.032	0.039	0.347	0.930	0.899
SE_2/RMSE	0.572	0.364	0.244	1.111	0.862	0.846
SE_3/RMSE	1.049	0.690	0.379	1.252	1.013	1.144

SE_1 = standard error from the information matrix
SE_2 = standard error from the Hessian matrix
SE_3 = standard error from the outer product matrix

Table 4.4: RMSE and Standard Error of MLE (T=60 σ^2=0.1)

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
λ_1 = **-2.0**						
RMSE	6.229	3.002	0.737	0.317	0.403	0.162
SE_1/RMSE	0.307	0.441	0.219	0.172	0.926	0.701
SE_2/RMSE	0.920	0.704	0.786	1.085	0.919	1.038
SE_3/RMSE	0.991	0.718	0.865	1.119	0.932	1.078
λ_1 = **-1.0**						
RMSE	4.844	2.284	0.555	0.170	0.379	0.163
SE_1/RMSE	0.398	0.580	0.297	0.216	0.984	0.704
SE_2/RMSE	0.953	0.757	0.853	1.036	0.977	1.002
SE_3/RMSE	1.094	0.804	0.987	1.212	1.046	1.138
λ_1 = **-0.5**						
RMSE	4.324	2.375	0.542	0.102	0.385	0.174
SE_1/RMSE	0.450	0.557	0.313	0.258	0.970	0.666
SE_2/RMSE	0.917	0.739	0.778	0.974	0.951	0.929
SE_3/RMSE	0.927	0.775	0.780	0.991	0.998	0.931
λ_1 = **0.1**						
RMSE	3.470	2.409	0.407	0.040	0.376	0.167
SE_1/RMSE	0.574	0.549	0.448	0.386	0.994	0.730
SE_2/RMSE	0.954	0.742	0.902	0.970	0.968	0.978
SE_3/RMSE	1.054	0.783	1.050	1.158	0.995	1.122
λ_1 = **0.5**						
RMSE	3.967	2.217	0.472	0.100	0.373	0.167
SE_1/RMSE	0.491	0.597	0.359	0.262	1.001	0.695
SE_2/RMSE	0.977	0.775	0.847	0.999	0.985	0.982
SE_3/RMSE	1.067	0.844	0.907	1.070	1.069	1.036
λ_1 = **1.0**						
RMSE	4.841	2.397	0.567	0.176	0.373	0.165
SE_1/RMSE	0.398	0.552	0.290	0.209	1.000	0.694
SE_2/RMSE	0.947	0.807	0.817	1.036	0.985	1.016
SE_3/RMSE	0.999	0.895	0.857	1.060	1.062	1.045
λ_1 = **2.0**						
RMSE	5.930	2.505	0.682	0.303	0.384	0.166
SE_1/RMSE	0.322	0.529	0.237	0.180	0.972	0.684
SE_2/RMSE	0.979	0.778	0.869	1.127	0.960	1.007
SE_3/RMSE	0.920	0.834	0.836	1.024	1.034	0.968

SE_1 = standard error from the information matrix
SE_2 = standard error from the Hessian matrix
SE_3 = standard error from the outer product matrix

Table 4.5: RMSE and Standard Error of MLE (T=60 σ^2=0.5)

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
λ_1 = **-2.0**						
RMSE	15.568	19.851	2.370	0.471	0.874	0.316
SE_1/RMSE	0.274	0.146	0.158	0.284	0.931	0.826
SE_2/RMSE	0.722	0.508	0.470	1.034	0.937	0.855
SE_3/RMSE	0.986	0.675	0.639	1.243	1.004	0.975
λ_1 = **-1.0**						
RMSE	15.400	26.234	1.674	0.266	0.907	0.308
SE_1/RMSE	0.284	0.113	0.224	0.338	0.920	0.836
SE_2/RMSE	0.686	0.420	0.548	1.020	0.911	0.951
SE_3/RMSE	0.892	0.542	0.770	1.184	0.958	1.089
λ_1 = **-0.5**						
RMSE	12.545	20.581	1.511	0.150	0.907	0.320
SE_1/RMSE	0.352	0.144	0.256	0.422	0.921	0.817
SE_2/RMSE	0.792	0.510	0.613	1.006	0.918	0.935
SE_3/RMSE	0.935	0.610	0.766	1.123	0.918	1.064
λ_1 = **0.1**						
RMSE	15.931	32.948	1.959	0.061	0.925	0.314
SE_1/RMSE	0.281	0.090	0.210	0.570	0.903	0.871
SE_2/RMSE	0.612	0.375	0.451	0.959	0.877	0.982
SE_3/RMSE	0.722	0.515	0.441	0.978	0.929	0.972
λ_1 = **0.5**						
RMSE	12.664	22.537	1.980	0.154	0.886	0.314
SE_1/RMSE	0.349	0.131	0.195	0.410	0.942	0.833
SE_2/RMSE	0.753	0.446	0.477	0.968	0.921	0.940
SE_3/RMSE	0.831	0.528	0.481	0.891	0.938	0.907
λ_1 = **1.0**						
RMSE	15.909	28.408	2.239	0.276	0.912	0.328
SE_1/RMSE	0.275	0.104	0.168	0.326	0.915	0.784
SE_2/RMSE	0.671	0.398	0.471	0.941	0.892	0.856
SE_3/RMSE	0.915	0.505	0.660	1.228	0.945	1.034
λ_1 = **2.0**						
RMSE	16.462	22.753	3.456	0.473	0.907	0.318
SE_1/RMSE	0.259	0.127	0.109	0.283	0.897	0.822
SE_2/RMSE	0.710	0.496	0.339	1.014	0.890	0.842
SE_3/RMSE	0.962	0.607	0.518	1.147	0.918	1.009

SE_1 = standard error from the information matrix
SE_2 = standard error from the Hessian matrix
SE_3 = standard error from the outer product matrix

Table 4.6: Kolmogorov-Smironov Statistics for Asymptotic χ^2_3-Distribution (σ^2=0.1)

	$Wald_1$	$Wald_2$	LM	LR
λ_1 = **-2.0**				
T=30	0.080	0.065	0.693	0.055
T=60	0.042	0.084	0.751	0.059
λ_1 = **-1.0**				
T=30	0.066	0.114	0.747	0.059
T=60	0.031	0.066	0.748	0.027
λ_1 = **-0.5**				
T=30	0.132	0.160	0.763	0.115
T=60	0.073	0.097	0.734	0.065
λ_1 = **0.1**				
T=30	0.153	0.150	0.842	0.117
T=60	0.044	0.058	0.685	0.029
λ_1 = **0.5**				
T=30	0.113	0.133	0.675	0.093
T=60	0.026	0.058	0.686	0.019
λ_1 = **1.0**				
T=30	0.077	0.142	0.692	0.060
T=60	0.033	0.056	0.771	0.033
λ_1 = **2.0**				
T=30	0.083	0.093	0.843	0.063
T=60	0.029	0.101	0.691	0.031

Table 4.7: Kolmogorov-Smironov Statistics for Asymptotic χ^2_3-Distribution (σ^2=0.5)

	Wald_1	Wald_2	LM	LR
λ_1 = **-2.0**				
T=30	0.195	0.089	0.756	0.018
T=60	0.125	0.053	0.669	0.017
λ_1 = **-1.0**				
T=30	0.201	0.116	0.768	0.073
T=60	0.084	0.066	0.700	0.038
λ_1 = **-0.5**				
T=30	0.149	0.160	0.793	0.082
T=60	0.068	0.097	0.718	0.041
λ_1 = **0.1**				
T=30	0.130	0.139	0.727	0.101
T=60	0.069	0.112	0.695	0.051
λ_1 = **0.5**				
T=30	0.165	0.168	0.711	0.102
T=60	0.067	0.174	0.724	0.033
λ_1 = **1.0**				
T=30	0.194	0.108	0.644	0.048
T=60	0.108	0.064	0.808	0.063
λ_1 = **2.0**				
T=30	0.191	0.109	0.711	0.039
T=60	0.134	0.114	0.707	0.034

Table 4.8: Real Size of the χ^2_3-Distribution
(T=30 σ^2=0.1)

	$Wald_1$	$Wald_2$	LM	LR
λ_1 = **-2.0**				
α=0.01	0.044	0.050	0.413	0.022
α=0.05	0.102	0.111	0.634	0.086
α=0.10	0.160	0.158	0.761	0.135
λ_1 = **-1.0**				
α=0.01	0.041	0.090	0.478	0.022
α=0.05	0.085	0.155	0.746	0.063
α=0.10	0.139	0.202	0.842	0.122
λ_1 = **-0.5**				
α=0.01	0.043	0.107	0.322	0.023
α=0.05	0.110	0.189	0.698	0.086
α=0.10	0.175	0.255	0.849	0.147
λ_1 = **0.1**				
α=0.01	0.057	0.105	0.650	0.030
α=0.05	0.127	0.180	0.860	0.093
α=0.10	0.208	0.244	0.938	0.163
λ_1 = **0.5**				
α=0.01	0.045	0.097	0.439	0.020
α=0.05	0.116	0.167	0.628	0.089
α=0.10	0.181	0.220	0.752	0.159
λ_1 = **1.0**				
α=0.01	0.039	0.094	0.300	0.015
α=0.05	0.097	0.163	0.629	0.075
α=0.10	0.153	0.236	0.771	0.126
λ_1 = **2.0**				
α=0.01	0.051	0.070	0.664	0.026
α=0.05	0.099	0.131	0.864	0.071
α=0.10	0.153	0.183	0.936	0.128

Table 4.9: Real Size of the χ^2_3-Distribution
(T=30 σ^2=0.5)

	Wald_1	Wald_2	LM	LR
λ_1 = **-2.0**				
α=0.01	0.196	0.068	0.472	0.015
α=0.05	0.243	0.133	0.757	0.056
α=0.10	0.288	0.177	0.852	0.109
λ_1 = **-1.0**				
α=0.01	0.194	0.088	0.479	0.014
α=0.05	0.250	0.162	0.762	0.068
α=0.10	0.297	0.213	0.865	0.124
λ_1 = **-0.5**				
α=0.01	0.116	0.123	0.371	0.027
α=0.05	0.182	0.197	0.734	0.065
α=0.10	0.238	0.255	0.871	0.137
λ_1 = **0.1**				
α=0.01	0.075	0.110	0.384	0.024
α=0.05	0.147	0.187	0.637	0.083
α=0.10	0.213	0.232	0.797	0.157
λ_1 = **0.5**				
α=0.01	0.133	0.124	0.300	0.022
α=0.05	0.202	0.203	0.603	0.074
α=0.10	0.250	0.268	0.764	0.142
λ_1 = **1.0**				
α=0.01	0.180	0.095	0.380	0.009
α=0.05	0.233	0.156	0.582	0.066
α=0.10	0.285	0.196	0.710	0.132
λ_1 = **2.0**				
α=0.01	0.198	0.070	0.320	0.009
α=0.05	0.239	0.143	0.645	0.053
α=0.10	0.275	0.200	0.779	0.097

Table 4.10: Real Size of the χ^2_3-Distribution
(T=60 σ^2=0.1)

	$Wald_1$	$Wald_2$	LM	LR
λ_1 = **-2.0**				
α=0.01	0.014	0.034	0.381	0.008
α=0.05	0.061	0.102	0.703	0.053
α=0.10	0.121	0.162	0.829	0.119
λ_1 = **-1.0**				
α=0.01	0.015	0.048	0.455	0.011
α=0.05	0.060	0.115	0.671	0.061
α=0.10	0.119	0.150	0.813	0.110
λ_1 = **-0.5**				
α=0.01	0.021	0.055	0.384	0.015
α=0.05	0.076	0.129	0.680	0.064
α=0.10	0.132	0.184	0.803	0.127
λ_1 = **0.1**				
α=0.01	0.023	0.051	0.345	0.016
α=0.05	0.081	0.103	0.613	0.067
α=0.10	0.127	0.149	0.745	0.122
λ_1 = **0.5**				
α=0.01	0.014	0.052	0.333	0.011
α=0.05	0.056	0.104	0.620	0.052
α=0.10	0.114	0.149	0.751	0.106
λ_1 = **1.0**				
α=0.01	0.016	0.039	0.554	0.010
α=0.05	0.049	0.101	0.795	0.052
α=0.10	0.110	0.153	0.870	0.110
λ_1 = **2.0**				
α=0.01	0.018	0.055	0.366	0.009
α=0.05	0.071	0.129	0.593	0.065
α=0.10	0.121	0.186	0.754	0.121

Table 4.11: Real Size of the χ^2_3-Distribution
(T=60 σ^2=0.5)

	$Wald_1$	$Wald_2$	LM	LR
λ_1 = **-2.0**				
α=0.01	0.118	0.036	0.325	0.007
α=0.05	0.172	0.189	0.573	0.061
α=0.10	0.215	0.147	0.714	0.110
λ_1 = **-1.0**				
α=0.01	0.082	0.056	0.504	0.007
α=0.05	0.132	0.113	0.698	0.055
α=0.10	0.174	0.152	0.797	0.119
λ_1 = **-0.5**				
α=0.01	0.058	0.058	0.367	0.010
α=0.05	0.115	0.138	0.683	0.059
α=0.10	0.159	0.194	0.806	0.117
λ_1 = **0.1**				
α=0.01	0.047	0.073	0.379	0.019
α=0.05	0.090	0.140	0.649	0.059
α=0.10	0.156	0.207	0.767	0.128
λ_1 = **0.5**				
α=0.01	0.048	0.098	0.253	0.009
α=0.05	0.092	0.186	0.616	0.054
α=0.10	0.152	0.255	0.793	0.105
λ_1 = **1.0**				
α=0.01	0.101	0.035	0.536	0.017
α=0.05	0.141	0.101	0.805	0.059
α=0.10	0.197	0.161	0.907	0.116
λ_1 = **2.0**				
α=0.01	0.139	0.060	0.370	0.014
α=0.05	0.181	0.131	0.648	0.058
α=0.10	0.226	0.193	0.780	0.113

Table 4.12: Asymptotic and Estimated Power

	$\sigma^2 = 0.1$				$\sigma^2 = 0.5$			
	$T = 30$		$T = 60$		$T = 30$		$T = 60$	
	$\mathcal{P}^{1)}$	$\hat{\mathcal{P}}^{2)}$	$\mathcal{P}$	$\hat{\mathcal{P}}$	$\mathcal{P}$	$\hat{\mathcal{P}}$	$\mathcal{P}$	$\hat{\mathcal{P}}$
Wald_1								
$\alpha = 0.01$	0.965	0.611	0.218	0.793	0.999	0.209	0.529	0.238
$\alpha = 0.05$	0.992	0.786	0.434	0.912	0.999	0.375	0.751	0.447
$\alpha = 0.10$	0.966	0.854	0.562	0.953	0.999	0.474	0.840	0.571
Wald_2								
$\alpha = 0.01$	0.965	0.290	0.218	0.621	0.999	0.051	0.529	0.117
$\alpha = 0.05$	0.992	0.474	0.434	0.799	0.999	0.131	0.751	0.257
$\alpha = 0.10$	0.966	0.575	0.562	0.872	0.999	0.211	0.840	0.358
LM								
$\alpha = 0.01$	0.965	0.867	0.218	0.910	0.999	0.470	0.529	0.694
$\alpha = 0.05$	0.992	0.962	0.434	0.998	0.999	0.715	0.751	0.897
$\alpha = 0.10$	0.966	0.983	0.562	1.000	0.999	0.835	0.840	0.959
LR								
$\alpha = 0.01$	0.965	0.520	0.218	0.758	0.999	0.141	0.529	0.209
$\alpha = 0.05$	0.992	0.731	0.434	0.905	0.999	0.324	0.751	0.428
$\alpha = 0.10$	0.966	0.823	0.562	0.951	0.999	0.432	0.840	0.555

1) $\mathcal{P}$ = asymptotic power
2) $\hat{\mathcal{P}}$ = estimated power

CHAPTER 5
SHRINKAGE ESTIMATION IN NONLINEAR REGRESSION: THE BOX-COX TRANSFORMATION

5.1 Introduction

In the normal linear regression model, the Stein-rule coefficient estimator is known to dominate the maximum likelihood (ML) estimator under quadratic loss for the case of more than two parameters if certain design related conditions are met. With the advent of increased computing power, the nonlinear regression model tends to be widely used in econometric applications. Therefore, it is worthwhile to analyze Stein-rule estimation within the framework of the nonlinear regression model. However, there have been only a few studies on Stein-like or biased estimation methods in the context of the nonlinear regression model [Schaefer, Roi and Wolfe (1984); Schaefer (1986); Adkins and Hill (1989)].

Schaefer, Roi and Wolfe (1984) proposed a ridge-type estimator for the logistic regression model and showed that the ridge estimator had smaller MSE than the ML estimator when the sample size was sufficiently large and the multicollinearity was severe. Schaefer (1986) employed a Stein-rule estimator in the logistic regression model. He showed that the Stein-rule estimator outperformed the ML estimator when the data were multicollinear.

In the context of the probit model, Adkins and Hill (1989) investigated the risk properties of a Stein-rule estimator, together with those of the constrained

ML and pretest estimators. The positive-part Stein-rule estimator outperformed the ML estimator and had smaller risk than the Stein-rule estimator for small to moderate degree of hypothesis error.

The objective of this Chapter is to propose a shrinkage (positive-part Stein-like) estimator for the Box-Cox model and to derive the asymptotic risk functions of the ML, constrained ML, pre-test and shrinkage estimators. The estimated (small sample) risk functions of these estimators will be examined using a Monte Carlo simulation.

In Section 5.2, a shrinkage estimator for the Box-Cox model will be presented, together with the ML, constrained ML, and pretest estimators. In Section 5.3, the asymptotic risks of the constrained ML, pretest, and shrinkage estimators will be derived and will be compared to the ML estimator's risk function. Section 5.4 will contain the design of simulation. The results will be discussed in Section 5.5 and conclusions presented in Section 5.6.

5.2 Alternative Estimators in the Box-Cox Regression Model

The general Box-Cox regression model is given by

$$\begin{aligned} T(y_t) &= \beta_1 + \beta_2 x_{2t}^{(\lambda_2)} + \cdots + \beta_k x_{kt}^{(\lambda_k)} + \epsilon_t \\ &= \mu_t + \epsilon_t,\ \epsilon_t \sim iidN(0,\sigma^2),\ t = 1,\ldots,T \end{aligned}$$

where $T(y_t)$ is a suitable transformation that makes the random variables y_t normally distributed and x_{it}, $i = 1, \ldots, k$, are assumed to be nonstochastic. Since the pre-transformed dependent variable y_t is truncated ($y_t > 0$), in order for the Box-Cox transformation to be well defined, the transformed variable $y_t^{(\lambda_1)}$ is also truncated [Hinkley (1975); Poirier (1978); Amemiya and Powell (1981)]:

$$y_t^{(\lambda_1)} = T(y_t) \quad \text{if } L < T(y_t) < R$$

where $R = -\frac{1}{\lambda_1}$ and $L = -\infty$ if $\lambda_1 < 0$; $R = +\infty$ and $L = -\frac{1}{\lambda_1}$ if $\lambda_1 > 0$. Let $\underline{\theta} = (\beta_1, \cdots, \beta_k, \lambda_1, \cdots, \lambda_k, \sigma^2)'$ and $K = dim(\underline{\theta})$. The log-likelihood function of the truncated Box-Cox model is given by

$$\begin{aligned} \ell(\underline{\theta}) &= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\epsilon_t^2 \\ &\quad - \sum_{t=1}^{T}\ln(\Phi(R_t) - \Phi(L_t)) + (\lambda_1 - 1)\sum_{t=1}^{T}\ln y_t \end{aligned} \tag{5.1}$$

where $R_t = (R - \mu_t)/\sigma$ and $L_t = (L - \mu_t)/\sigma$ and $\epsilon_t = y_t^{(\lambda_1)} - \mu_t$.

The ML estimator $\hat{\underline{\theta}}$ is obtained by solving the first-order conditions for the log-likelihood function (5.1). Consider the linear restriction

$$H_0 : R\underline{\theta} - \underline{r} = \underline{0}$$

where R is a $J \times K$ matrix of rank J.

The constrained ML estimator $\tilde{\underline{\theta}}$ is obtained by maximizing the Lagrangean function

$$M(\underline{\theta}, \underline{\mu}) = \ell(\underline{\theta}) - \underline{\mu}'(R\underline{\theta} - \underline{r}) \tag{5.2}$$

where $\underline{\mu}$ is a $J \times 1$ Lagrange multiplier vector. Then the pre-test estimator $\underline{\theta}^{\star}$ is defined as

$$\underline{\theta}^{\star} = \begin{cases} \tilde{\underline{\theta}} & \text{if } u \leq C_{\alpha} \\ \hat{\underline{\theta}} & \text{if } u > C_{\alpha} \end{cases}$$

where u is an appropriate $\chi^2_{(J)}$-statistic (e.g., Lagrange multiplier (LM), likelihood ratio (LR) or Wald test) for testing $H_0 : R\underline{\theta} = \underline{r}$, and C_{α} is the critical value of $\chi^2_{(J)}$-distribution at the significance level α. Using indicator functions, the pretest estimator can be written

$$\underline{\theta}^{\star} = I_{(0,C_{\alpha}]}(u)\tilde{\underline{\theta}} + I_{(C_{\alpha},\infty)}(u)\hat{\underline{\theta}}$$

Finally, the shrinkage (positive-part Stein-like) estimator proposed is:

$$\underline{\theta}^{+} = I_{(a,\infty)}(u)(1 - \frac{a}{u})(\hat{\underline{\theta}} - \tilde{\underline{\theta}}) + \tilde{\underline{\theta}}$$

where a is a constant such that $a > J$. Since the critical value C_{α} at the usual significance level ($\alpha = 0.01, 0.05$ and 0.10) exceeds the degrees of freedom J of the $\chi^2_{(J)}$-distribution, C_{α} will be selected as the value of a. When the test statistic is less than the value of a, the shrinkage estimator is reduced to the constrained ML estimator while the linear combination of the ML and constrained ML estimators will be obtained if the statistic u exceeds the value of a. The more confidence in the null hypothesis we have, the larger the value of a will be selected. In contrast, the positive-part Stein-rule estimator in the linear regression model takes the value of a on a certain closed interval for minimaxity. Our shrinkage estimator tends to put a heavier weight on the null hypothesis than the positive-part Stein-rule

estimator in the linear regression model since the asymptotic risk gain is obtained under the null hypothesis. Let $\omega = 1 - I_{(a,\infty)}(u)(1 - \frac{a}{u})$. The shrinkage estimator $\underline{\theta}^+$ can also be written as a linear combination of the ML and constrained ML estimators:

$$\underline{\theta}^+ = \omega\tilde{\underline{\theta}} + (1-\omega)\hat{\underline{\theta}}$$

The shrinkage estimator has the property of shrinking every element of the ML estimator $\hat{\underline{\theta}}$ toward the corresponding element of the constrained ML estimator $\tilde{\underline{\theta}}$. Therefore, the Stein-like estimator may have significant risk gain if the constrained ML estimator $\tilde{\underline{\theta}}$ is close to the true parameter vector $\underline{\theta}_0$.

5.3 Asymptotic Risk Properties

Before we derive the asymptotic risk function of the Stein-like estimator, several notations and relevant assumptions are given. We denote the score vector and Hessian matrix as follows:

$$\begin{aligned} \underline{q}(\underline{\theta}) &= \frac{\partial \ell}{\partial \underline{\theta}} \\ Q(\underline{\theta}) &= \frac{\partial^2 \ell}{\partial \underline{\theta} \partial \underline{\theta}'} \\ Q_T(\underline{\theta}) &= \frac{1}{T} Q(\underline{\theta}) \end{aligned}$$

In addition, the information matrix is represented by $\Im(\underline{\theta}_0) = -EQ(\underline{\theta}_0)$ and $\Im_T(\underline{\theta}_0) = \frac{1}{T}\Im(\underline{\theta}_0)$. The dependence of the score vector and information matrix on the parameter vector will often be omitted, e.g., the score vector will be denoted by $\underline{q}$, rather than $\underline{q}(\underline{\theta}_0)$. We want to make the following assumptions on the

score vector and information matrix:

1. $[\Im]_{ij} = O(T)$ as $T \to \infty$ for $i, j = 1, \ldots, K$. $\lim_{T\to\infty} \Im_T = \Im_\infty$ exists and is a non-singular matrix.

2. $\text{plim}\Im_T^{-1}(-Q_T) = I_K$ and $\text{plim}Q_T^{-1}(\underline{\theta}_0)Q_T(\underline{\theta}^*) = I_K$ for $\underline{\theta}^*$ which is contained in the neighborhood of $\underline{\theta}_0$.

3. $\Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q} \overset{a}{\sim} N(\underline{0}, I_K)$

Definition 5.1 *Let $\{X_T\}$ be a sequence of random variables with distribution functions $\{F_T\}$ and X be a random variable with the distribution function $\{F\}$. Suppose $X_T \xrightarrow{d} X$. Then the asymptotic expectation of X_T is defined by*

$$\mathcal{A}E(X_T) = E_F(X)$$

For the measurement of estimator's risk, the asymptotic risk function is defined as

$$\varrho(\underline{\bar{\theta}}, \underline{\theta}_0) = \mathcal{A}E[T(\underline{\bar{\theta}} - \underline{\theta}_0)'\Im_T(\underline{\bar{\theta}} - \underline{\theta}_0)]$$

where $\underline{\bar{\theta}}$ is the estimator of $\underline{\theta}_0$ such that $X_T = T(\underline{\bar{\theta}} - \underline{\theta}_0)'\Im_T(\underline{\bar{\theta}} - \underline{\theta}_0)$ converges in distribution to a random variable X.

Consider the asymptotic relationship

$$\Im_T^{1/2}\sqrt{T}(\underline{\hat{\theta}} - \underline{\theta}_0) \overset{A}{=} \Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q}$$

where $\overset{A}{=}$ represents asymptotic equivalence. Since $\Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q} \overset{a}{\sim} N(\underline{0}, I_K)$, we can infer that

$$\Im_T^{1/2}\sqrt{T}(\underline{\hat{\theta}} - \underline{\theta}_0) \xrightarrow{d} \underline{z}$$

where $\underline{z} \sim N(\underline{0}, I_K)$. Thus

$$\frac{1}{\sqrt{T}}\underline{q} \stackrel{A}{=} \Im_T^{1/2}\underline{z} \tag{5.3}$$

and

$$\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) \stackrel{A}{=} \Im_T^{-1/2}\underline{z} \tag{5.4}$$

The elements of the constrained ML estimator $\tilde{\underline{\theta}}$ satisfy the first-order conditions for the Lagrangean function (5.2):

$$\tilde{\underline{q}} - R'\tilde{\underline{\mu}} = \underline{0}$$

$$R\tilde{\underline{\theta}} - \underline{r} = \underline{0}$$

By taking a first order Taylor series expansion of $\tilde{\underline{q}}$ about $\underline{\theta}_0$,

$$\begin{aligned}\tilde{\underline{q}} &\stackrel{A}{=} \underline{q} + Q(\tilde{\underline{\theta}} - \underline{\theta}_0) \\ &\stackrel{A}{=} \underline{q} - \Im(\tilde{\underline{\theta}} - \underline{\theta}_0)\end{aligned}$$

Therefore, under the null hypothesis H_0,

$$-\Im_T\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) + \frac{1}{\sqrt{T}}\underline{q} - \frac{1}{\sqrt{T}}R'\tilde{\underline{\mu}} \stackrel{A}{=} \underline{0} \tag{5.5}$$

$$\sqrt{T}R(\tilde{\underline{\theta}} - \underline{\theta}_0) = \underline{0} \tag{5.6}$$

Using matrix notation, (5.5) and (5.6) are written as

$$\begin{bmatrix} \Im_T & R' \\ R & 0_{J\times J} \end{bmatrix} \begin{bmatrix} \sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) \\ \frac{1}{\sqrt{T}}\tilde{\underline{\mu}} \end{bmatrix} \stackrel{A}{=} \begin{bmatrix} \frac{1}{\sqrt{T}}\underline{q} \\ \underline{0} \end{bmatrix}$$

Then, we obtain the following relationship:

$$\begin{bmatrix} \sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) \\ \frac{1}{\sqrt{T}}\tilde{\underline{\mu}} \end{bmatrix} \stackrel{A}{=} \begin{bmatrix} \Im_T & R' \\ R & 0_{J\times J} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\sqrt{T}}\underline{q} \\ \underline{0} \end{bmatrix}$$

Consequently,

$$\begin{aligned}\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) &\stackrel{A}{=} \Im_T^{-1}[I_K - R'(R\Im_T^{-1}R')^{-1}R\Im_T^{-1}]\frac{1}{\sqrt{T}}\underline{q} \\ &= \Im_T^{-1/2}[I_K - \Im_T^{-1/2}R'(R\Im_T^{-1}R')^{-1}R\Im_T^{-1/2}]\Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q} \\ &= \Im_T^{-1/2}[I_K - W]\Im_T^{-1/2}\frac{1}{\sqrt{T}}\underline{q}\end{aligned}$$

where $W = \Im_T^{-1/2}R'(R\Im_T^{-1}R')^{-1}R\Im_T^{-1/2}$ is an idempotent and symmetric matrix of rank J. Therefore, the following asymptotic equivalance is obtained:

$$\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) \stackrel{A}{=} \Im_T^{-1/2}[I_K - W]\underline{z} \tag{5.7}$$

Proposition 5.1 *The ML estimator $\hat{\underline{\theta}}$ has the asymptotic risk function $\varrho(\hat{\underline{\theta}}, \underline{\theta}_0) = K$ under the weighted quadratic loss $\mathcal{L} = T(\hat{\underline{\theta}} - \underline{\theta}_0)'\Im_T(\hat{\underline{\theta}} - \underline{\theta}_0)$.*

Proof:

From (5.3),

$$\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) \stackrel{A}{=} \Im_T^{-1/2}\underline{z}$$

Then the loss function is asymptotically

$$\mathcal{L} \stackrel{A}{=} \underline{z}'\underline{z} \sim \chi^2_{(K)}$$

since $\underline{z} \sim N(\underline{0}, I_K)$. Therefore,

$$\varrho(\hat{\underline{\theta}}, \underline{\theta}_0) = K$$

Proposition 5.2 *The constrained ML estimator $\tilde{\underline{\theta}}$ has the asymptotic risk function $\varrho(\tilde{\underline{\theta}}, \underline{\theta}_0) = K - J$ under the weighted quadratic loss $\mathcal{L} = T(\tilde{\underline{\theta}} - \underline{\theta}_0)'\Im_T(\tilde{\underline{\theta}} - \underline{\theta}_0)$ if $H_0 : R\underline{\theta} = \underline{r}$ is true.*

Proof:

From (5.7),

$$\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) \stackrel{A}{=} \Im_T^{-1/2}(I_K - W)\underline{z}$$

Then the loss function is asymptotically

$$\mathcal{L} \stackrel{A}{=} \underline{z}'(I_K - W)\underline{z} \sim \chi^2_{(K-J)}$$

since $\underline{z} \sim N(\underline{0}, I_K)$ *and* $I_K - W$ *is a symmetric and idempotent matrix of rank* $K - J$. *Therefore,*

$$\varrho(\tilde{\underline{\theta}}, \underline{\theta}_0) = K - J$$

Proposition 5.3 *The* α*-level pretest estimator* $\underline{\theta}^*$ *has the asymptotic risk function* $\varrho(\underline{\theta}^*, \underline{\theta}_0) = K - J(1 - \int_{C_\alpha}^{\infty} d\chi^2_{(J+2)})$ *under the weighted quadratic loss* $\mathcal{L} = T(\underline{\theta}^* - \underline{\theta}_0)'\Im_T(\underline{\theta}^* - \underline{\theta}_0)$ *if* $H_0 : R\underline{\theta} = \underline{r}$ *is true.*

Proof:

Consider the following pretest estimator

$$I_{(0,C_\alpha]}(u)\tilde{\underline{\theta}} + (1 - I_{(0,C_\alpha]}(u))\hat{\underline{\theta}}$$

From (5.4) *and* (5.7),

$$\begin{aligned}
\sqrt{T}(\underline{\theta}^* - \underline{\theta}_0) &= I_{(0,C_\alpha]}(u)\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) + (1 - I_{(0,C_\alpha]}(u))\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) \\
&\stackrel{A}{=} I_{(0,C_\alpha]}(u)\Im_T^{-1/2}(I_K - W)\underline{z} + (1 - I_{(0,C_\alpha]}(u))\Im_T^{-1/2}\underline{z} \\
&= (1 - I_{(0,C_\alpha]}(u))\Im_T^{-1/2}W\underline{z} + \Im_T^{-1/2}(I_K - W)\underline{z}
\end{aligned}$$

The loss function is asymptotically

$$\begin{aligned}\mathcal{L} &\stackrel{A}{=} \underline{z}'[(1 - I_{(0,C_\alpha]}(u))W + (I_K - W)]\underline{z} \\ &= I_{(C_\alpha,\infty)}(u)\underline{z}'W\underline{z} + \underline{z}'(I_K - W)\underline{z}\end{aligned}$$

The asymptotic risk function of the α-level pretest estimator is written as

$$\begin{aligned}\varrho(\underline{\theta}^*, \underline{\theta}_0) &= EI_{(C_\alpha,\infty)}(u)\underline{z}'W\underline{z} + E\underline{z}'(I_K - W)\underline{z} \\ &= EI_{(C_\alpha,\infty)}(u)u + (K - J)\end{aligned}$$

since $u = \underline{z}'W\underline{z} \sim \chi^2_{(J)}$ and $\underline{z}'(I_K - W)\underline{z} \sim \chi^2_{(K-J)}$. Therefore,

$$\begin{aligned}\varrho(\underline{\theta}^*, \underline{\theta}_0) &= \int_{C_\alpha}^{\infty} \frac{1}{\Gamma(J/2)2^{J/2}} u^{J/2} e^{-u/2} du + (K - J) \\ &= \int_{C_\alpha}^{\infty} \frac{J}{\Gamma((J+2)/2)2^{(J+2)/2}} u^{(J+2)/2-1} e^{-u/2} du + (K - J) \\ &= J \int_{C_\alpha}^{\infty} d\chi^2_{(J+2)} + (K - J) \\ &= K - J(1 - \int_{C_\alpha}^{\infty} d\chi^2_{(J+2)})\end{aligned}$$

Proposition 5.4 *The shrinkage estimator $\underline{\theta}^+$ has the asymptotic risk function $\varrho(\underline{\theta}^+, \underline{\theta}_0) = \frac{1}{\Gamma(J/2)} e^{-a/2}(a/2)^{J/2-2} + K - J$ under the weighted quadratic loss $\mathcal{L} = T(\underline{\theta}^+ - \underline{\theta}_0)'\Im_T(\underline{\theta}^+ - \underline{\theta}_0)$ if the null hypothesis $H_0 : R\underline{\theta} = \underline{r}$ holds and $a > J$.*

Proof:

$$\begin{aligned}\sqrt{T}(\underline{\theta}^+ - \underline{\theta}_0) &= \omega\sqrt{T}(\tilde{\underline{\theta}} - \underline{\theta}_0) + (1 - \omega)\sqrt{T}(\hat{\underline{\theta}} - \underline{\theta}_0) \\ &\stackrel{A}{=} \omega\Im_T^{-1/2}(I_K - W)\underline{z} + (1 - \omega)\Im_T^{-1/2}\underline{z} \\ &= (1 - \omega)\Im_T^{-1/2}W\underline{z} + \Im_T^{-1/2}(I_K - W)\underline{z}\end{aligned}$$

Then the loss function is asymptotically

$$\begin{aligned}\mathcal{L} &\stackrel{A}{=} \underline{z}'[(1-\omega)^2 W + (I_K - W)]\underline{z} \\ &= (1-\omega)^2 \underline{z}' W \underline{z} + \underline{z}'(I_K - W)\underline{z}\end{aligned}$$

The asymptotic risk function is written as

$$\begin{aligned}\varrho(\underline{\theta}^+, \underline{\theta}_0) &= E(1-\omega)^2 \underline{z}' W \underline{z} + E\underline{z}'(I_K - W)\underline{z} \\ &= EI_{(a,\infty)}(u)(1-\frac{a}{u})^2 u + K - J\end{aligned}$$

since $u = \underline{z}' W \underline{z} \sim \chi^2_{(J)}$ *and* $\underline{z}'(I_K - W)\underline{z} \sim \chi^2_{(K-J)}$. *Consider*

$$\begin{aligned}EI_{(a,\infty)}(u)(1-\frac{a}{u})^2 u &= \int_a^\infty (1-\frac{a}{u})^2 u \frac{1}{\Gamma(J/2)2^{J/2}} u^{J/2-1} e^{-u/2} du \\ &= \int_a^\infty \frac{1}{\Gamma(J/2)2^{J/2}} u^{J/2} e^{-u/2} du - 2a \int_a^\infty \frac{1}{\Gamma(J/2)2^{J/2}} u^{J/2-1} e^{-u/2} du \\ &\quad + a^2 \int_a^\infty \frac{1}{\Gamma(J/2)2^{J/2}} u^{J/2-2} e^{-u/2} du \\ &= \frac{2}{\Gamma(J/2)} \int_{a/2}^\infty x^{J/2} e^{-x} dx - \frac{2a}{\Gamma(J/2)} \int_{a/2}^\infty x^{J/2-1} e^{-x} dx \\ &\quad + \frac{a^2}{2\Gamma(J/2)} \int_{a/2}^\infty x^{J/2-2} e^{-x} dx\end{aligned}$$

Note that

$$\int_a^\infty \frac{1}{\Gamma(J)} x^{J-1} e^{-x} dx = \sum_{m=0}^{J-1} \frac{a^m e^{-a}}{m}$$

and $a > J$. *Using the approximation*

$$\int_a^\infty x^{J-1} e^{-x} dx \simeq e^{-a} a^{J-1}[1 + (J-1)/a + (J-1)(J-2)/a^2] \text{ [Maasoumi (1978)]}$$

we obtain

$$EI_{(a,\infty)}(u)(1-\frac{a}{u})^2 u \simeq \frac{1}{\Gamma(J/2)} e^{-a/2} (a/2)^{J/2-2}$$

Consequently, the risk function becomes

$$\varrho(\underline{\theta}^+, \underline{\theta}_0) = \frac{1}{\Gamma(J/2)} e^{-a/2}(a/2)^{J/2-2} + K - J$$

Proposition 5.5 *For any $a > J$, the asymptotic risk of the shrinkage estimator $\underline{\theta}^+$ dominates that of the ML estimator $\hat{\underline{\theta}}$ under the weighted quadratic loss $\mathcal{L} = T(\underline{\theta}^+ - \underline{\theta}_0)'\Im_T(\underline{\theta}^+ - \underline{\theta}_0)$ if the null hypothesis $H_0 : R\underline{\theta} = \underline{r}$ holds and $J \geq 2$.*

Proof:

1) *If $J > 4$*

$$\varrho(\underline{\theta}^+, \underline{\theta}_0) = \frac{1}{\Gamma(J/2)} e^{-a/2}(a/2)^{J/2-2} + K - J$$

Consider the first and second derivatives of the risk function with respect to a:

$$\begin{aligned} \frac{\partial \varrho}{\partial a} &= \frac{1}{2\Gamma(J/2)} e^{-a/2}(\frac{a}{2})^{J/2-3}[(J/2-2) - (a/2)] \\ \frac{\partial^2 \varrho}{\partial a^2} &= \frac{1}{16\Gamma(J/2)} e^{-a/2}(\frac{a}{2})^{J/2-4}[(a-J+4)^2 - 2(J-4)] \end{aligned}$$

The solutions for the first-order condition are $a = J-4$ and $a = 0$. Since $\frac{\partial^2 \varrho}{\partial a^2}|_{a=J-4} < 0$, the risk function has a maximum at $a = J-4$ for $a \geq 0$. The maximum risk is

$$\varrho_{max} = \frac{1}{\Gamma(J/2)} e^{-J/2+2}(J/2-2)^{J/2-2} + K - J$$

Using Stirling's formula $\Gamma(J-1) \simeq \sqrt{2\pi(J-2)}(\frac{J/2-2}{e})^{J/2-2}$, we obtain

$$\varrho_{max} \simeq \frac{1}{(J/2-1)\sqrt{2\pi(J/2-2)}} + K - J$$

which is less than K. Therefore, for any $a > J$,

$$\varrho(\tilde{\underline{\theta}}, \underline{\theta}_0) < K$$

since ϱ *is monotonically decreasing if* $a \geq J - 4$.

2) *If* $2 \leq J \leq 4$

Since $a > J$,

$$\frac{1}{\Gamma(J/2)} e^{-a/2}(a/2)^{J/2-2} < 1 \ \textit{for}\ J = 2, 3, 4$$

Therefore,

$$\varrho(\tilde{\underline{\theta}}, \underline{\theta}_0) < K$$

Under the null hypothesis, the constrained ML estimator has the smallest risk function. The pretest estimator and shrinkage estimator have risk between those of the ML and constrained ML estimators. The risk of the pretest estimator increases as the significance level declines. The greater the value of the constant a, the larger the risk gain of the shrinkage estimator can we obtain since the risk function is monotonically decreasing with respect to a.

5.4 Design of Monte Carlo Experiments

For our analysis, the data were generated by the model:

$$y_t^{(\lambda_1)} = 10.0 + 1.5x_{2t}^{(0.1)} - 0.5x_{3t}^{(1.0)} + \epsilon_t$$

where $\lambda_1 = (0.1, 0.5, 1.0, 2.0)$ and $\sigma^2 = (0.1, 0.5)$. The error disturbances ϵ_t were obtained using the GAUSS $N(0,1)$ random number generator RNDNS such that $y_t^{(\lambda_1)} > -\frac{1}{\lambda_1}$ since we can only observe truncated values of y_t ($y_t > 0$). The model was chosen to have two explanatory variables whose values were generated from linear combinations of uniform random numbers (GAUSS function RNDUS) such that $corr(x_{2t}, x_{3t}) = 0.2$:

$$\begin{aligned} x_{2t} &= 12 + 4u_t \\ x_{3t} &= 15 + u_t + 2\sqrt{6}v_t \end{aligned}$$

where u_t and v_t are uniformly distributed on the interval $U(-\sqrt{3}, \sqrt{3})$. The variances of the error term ($\sigma^2 = 0.1$ and 0.5) account for approximately 2 % and 6 % of the variation of the right-hand side of the Box-Cox model, respectively. Each Box-Cox model was estimated 1000 times for samples of size 30 and 60. We considered the null hypothesis $H_0 : \lambda_1 = \lambda_2 = 0$ and $\lambda_3 = 1.0$. Let $\bar{\theta}_{ij}$ be the ith parameter estimator in the jth replication of simulation. The estimated risk and MSE are defined by

$$\begin{aligned} \text{Risk under Quadratic Loss} &= \sum_{j=1}^{N} (\underline{\bar{\theta}}_j - \underline{\theta})'(\underline{\bar{\theta}}_j - \underline{\theta})/N \\ \text{Risk under Weighted Quadratic Loss} &= \sum_{j=1}^{N} (\underline{\bar{\theta}}_j - \underline{\theta})'\Im^{-1}(\underline{\bar{\theta}}_j - \underline{\theta})/N \\ \text{MSE of an Individual Parameter Estimator} &= \sum_{j=1}^{N} (\hat{\theta}_{ij} - \theta_i)^2/N \end{aligned}$$

where θ_i is the ith parameter of the parameter vector $\underline{\theta}$; $\underline{\bar{\theta}}_j$ is the estimator for $\underline{\theta}$ in the jth replication of simulation; and N represents the number of simulation

replicates. The information matrix $\Im$ was obtained using numerical integration (GAUSS function INTQUAD1). For testing the null hypothesis, we employed the likelihood ratio test since it uses the information from the unconstrained ML and constrained ML estimators.

5.5 Results

5.5.1 Risk Properties in the Model $\sigma^2 = 0.1$

When we define the hypothesis error as the distance of the hypothesized value from the true parameter value, the size of hypothesis error is proportional to the value of λ_1. Results for the estimated risks under quadratic loss for $\sigma^2 = 0.1$ are reported in Table 5.1. The estimated risks for all estimators decrease when the hypothesis error is reduced ($\lambda_1 = 2.0 \rightarrow \lambda_1 = 0.1$). For relatively small hypothesis error ($\lambda_1 = 0.1$) at $T = 30$, the ML estimator has the largest risk under quadratic loss. When the hypothesis error increases, the pretest estimators have the same value of risk as the ML estimator since all the LR tests reject the null hypothesis. The pretest estimators at 0.01 and 0.05 levels of significance have higher risk than the ML estimator for the model $\lambda_1 = 0.5$ at $T = 30$. The pretest estimators at all significance levels display higher risk than the ML estimator for the model $\lambda_1 = 1$ when $T = 60$. The shrinkage estimator dominates the ML estimator and the pretest estimator over the whole range of hypothesis error. According to our analytical results (Section 5.3), the risk of the pretest estimator is expected to be smaller at the higher significance level than at the lower significance level.

Analytically, the asymptotic risk of the constrained ML estimator is less than that of the ML estimator under the null hypothesis. If we take the conservative standard (high significance level), we are more likely to choose the constrained ML estimator (which has small risk values) than the unconstrained ML estimator (which leads to small risk of pretesting). However, the risk of the pretest estimator declines with the lower significance level (larger value of α) when the risk of the ML estimator is less than that of the pretest estimator (for $\lambda_1 = 0.1$ at $T = 60$; for $\lambda_1 = 0.5$ at $T = 30$).

For the models $\lambda_1 = 0.1$ and $\lambda_1 = 0.5$, the constrained ML estimator dominates other (ML, pretest and shrinkage) estimators at $T = 30$. The shrinkage estimator has smaller risk values than the ML estimator, the constrained ML estimator and the pretest estimator for the models λ_1=0.5 and 1.0 (T=30 and 60). In addition, the shrinkage estimator dominates the ML estimator over the whole space of hypothesis error for all sample sizes. Asymptotically, the shrinkage estimator using the larger critical value has smaller risk than the estimator using the smaller critical value. The results of our experiments are shown to correspond to the asymptotic results. When the sample size increases, all estimators but the constrained estimator have reduced risk. If $\lambda_1 \geq 0.5$, the risk of the ML estimator is lower than that of the constrained ML estimator.

When we consider the risk gain over the ML estimator (Tables 5.5 and 5.7), the shrinkage estimator performs well. For the model with small hypothesis error

($\lambda_1 = 0.1$), the risk of the shrinkage estimator is less than 50 % of risk of the ML estimator at $T = 30$. The risk gain of the constrained ML estimator is similar to that of the shrinkage estimator for large hypothesis error. However, the shrinkage estimator dominates all other estimators at $T = 60$.

The MSE's for $\sigma^2 = 0.1$ are reported in Tables 5.9–5.12. The MSE's of the estimators for β_1 and β_2 account for almost 98% of the estimated risk under quadratic loss. We observe that the MSE of the estimator for β_2 in constrained ML estimation is very small relative to the MSE's for other (ML, pretest and shrinkage) estimation methods. Therefore, the low risk for the constrained ML estimator can be explained by the remarkably small MSE of the estimator for β_2. All MSE's except for the constrained ML estimator are reduced when the sample has more observations. The magnitude of reduction in the MSE of the estimator for β_2 is conspicuous with increasing sample size.

The estimated risks (Table 5.3) under weighted quadratic loss have similar properties to those under quadratic loss, but the magnitude of the risks under weighted quadratic loss is much larger than under quadratic loss.

5.5.2 Risk Properties in the Model $\sigma^2 = 0.5$

The estimated risks for the model $\sigma^2 = 0.5$ are given in Tables 5.2 and 5.4. For all models (λ_1=0.1, 0.5, 1.0 and 2.0), the shrinkage estimator dominates the ML estimator. It is worth mentioning that the risk of the constrained ML estimator is much lower than that of any other estimator over the whole range of hypothesis

error, regardless of the loss function. When the error variance is relatively large ($\sigma^2 = 0.5$), the estimated risk for $\lambda_1 = 2.0$ is smaller than that for $\lambda_1 = 1.0$. From tables 5.13–5.16, we can observe that the MSE's of the linear parameter estimators for the model $\lambda_1 = 2.0$ are more stable than for the model $\lambda_1 = 1.0$. With increasing sample size, the risks are reduced except for the constrained ML estimator. The pretest estimators for all models have lower risk than the ML estimators at $T = 30$. However, the risks of the pretest estimators are higher than those of the ML estimators for $\lambda_1 = 0.1$ and $\lambda_1 = 2.0$ when $T = 60$. The risk of the pretest estimator is higher with a lower significance level when the ML estimator has higher risk than the pretest estimator. If the pretest estimator has higher risk than the ML estimator, the reverse is true.

The estimated MSE's for the model $\sigma^2 = 0.5$ are reported in Tables 5.13–5.16. The constrained ML estimators show a conspicuously small value for estimated risk, which may be explained by the observation that the MSE's of the linear parameter estimators in the constrained ML estimation are very small, even though the hypothesis error is substantial (e.g., $\lambda_1 = 2.0$). The risk gain of the shrinkage estimator and the constrained ML estimator over the ML estimator is larger when the error variance increases (Tables 5.5–5.8). The constrained ML estimator, especially, has very low risk ratio relative to the ML estimator even though the hypothesis error is large. Generally, the MSE's are reduced when the sample has more observation, except in the case of the constrained ML estimators.

The magnitude of reduction in the MSE of the estimator for β_2 is particularly substantial.

5.6 Conclusions

We can summarize the results as follows:

1. According to our analytical results, the shrinkage estimator dominates the ML estimator. The results of our simulation experiments are in accord with the asymptotic results in small samples.

2. The risk of the pretest estimator is less at higher significance levels than at lower levels of significance. The reverse is true when the ML estimator has lower risk than the pretest estimator.

3. The risk of the ML estimator is higher than that of the constrained ML estimator for the model $\sigma^2 = 0.1$ at $T = 30$. However, the ML estimator has lower risk than the constrained ML estimator as hypothesis error increases for the model $\sigma^2 = 0.1$ at $T = 60$.

4. Asymptotically, the risk function of the shrinkage estimator is monotonically decreasing with respect to the constant a. Our Monte Carlo experiments support this asymptotic result.

5. The risks and MSE's for the ML, pretest and shrinkage estimators decrease when the sample size increases. The risk of the constrained ML estimator

increases as the sample size becomes larger.

6. When the variance of the error disturbances is large ($\sigma^2 = 0.5$), the constrained ML estimators dominate the ML estimators over the entire range of hypothesis error. The estimated risk for the model $\lambda_1 = 2.0$ is smaller than for the model $\lambda_1 = 1.0$ when the variance of the error term is 0.5.

7. The risk gain of the shrinkage estimators over the ML estimators increases when the variance of the error disturbances becomes larger.

Table 5.1: Estimated Risk under Quadratic Loss
($\sigma^2 = 0.1$)

	$\lambda_1 = 0.1$	$\lambda_1 = 0.5$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
ML				
T=30	70.018	81.110	99.908	155.478
T=60	18.182	21.060	29.707	42.188
RML				
T=30	10.968	43.547	61.136	78.337
T=60	11.734	44.412	61.780	78.714
Pretest^1				
T=30	62.823	81.339	99.908	155.478
T=60	19.161	21.060	29.707	42.188
Pretest^2				
T=30	68.631	81.153	99.908	155.478
T=60	18.636	21.060	29.707	42.188
Pretest^3				
T=30	69.296	81.110	99.908	155.478
T=60	18.394	21.060	29.707	42.188
Stein^1				
T=30	18.508	43.040	58.050	91.697
T=60	9.689	16.379	23.049	29.841
Stein^2				
T=30	26.500	51.262	67.435	107.401
T=60	10.356	17.023	24.375	32.756
Stein^3				
T=30	31.714	55.931	72.632	115.537
T=60	11.049	17.542	25.176	34.312

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.2: Estimated Risk under Quadratic Loss
($\sigma^2 = 0.5$)

	$\lambda_1 = 0.1$	$\lambda_1 = 0.5$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
ML				
T=30	11824.120	19789.498	60510.114	16830.985
T=60	1344.336	673.681	1073.080	806.540
RML				
T=30	11.170	43.860	61.315	78.605
T=60	12.045	44.754	62.085	78.946
Pretest^1				
T=30	5366.218	18620.618	60099.594	16718.281
T=60	954.334	674.051	1074.974	807.957
Pretest^2				
T=30	10079.830	19616.519	60439.262	16822.778
T=60	1267.982	674.045	1073.721	807.041
Pretest^3				
T=30	10942.179	19715.697	60480.799	16830.254
T=60	1298.441	673.599	1073.234	806.687
Stein^1				
T=30	258.984	5240.702	13751.621	4776.065
T=60	132.837	334.683	616.080	431.872
Stein^2				
T=30	1094.451	8270.790	24468.786	7589.888
T=60	278.965	414.950	735.739	530.328
Stein^3				
T=30	1959.614	10008.052	30299.999	9101.661
T=60	401.094	457.328	794.986	578.983

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.3: Estimated Risk under Weighted Quadratic Loss
($\sigma^2 = 0.1$)

	$\lambda_1 = 0.1$	$\lambda_1 = 0.5$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
ML				
T=30	44392.379	61633.355	72920.940	92273.179
T=60	8209.787	8445.625	10372.345	13746.246
RML				
T=30	59.756	2386.251	7832.190	20067.569
T=60	135.644	5591.552	18114.402	45910.741
Pretest^1				
T=30	39447.556	61652.564	72920.940	92273.179
T=60	7933.022	8445.625	10372.345	13746.246
Pretest^2				
T=30	43393.983	61637.736	72920.940	92273.179
T=60	8144.091	8445.625	10372.345	13746.246
Pretest^3				
T=30	43971.873	61633.355	72920.940	92273.179
T=60	8177.266	8445.625	10372.345	13746.246
Stein^1				
T=30	8270.496	28804.231	39526.258	57198.504
T=60	2168.109	5806.984	8468.237	12009.737
Stein^2				
T=30	15129.336	37492.584	48445.415	66588.897
T=60	3444.054	6469.616	8839.795	12146.497
Stein^3				
T=30	19260.412	41771.774	52812.192	71178.822
T=60	4173.019	6808.171	9067.696	12323.234

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.4: Estimated Risk under Weighted Quadratic Loss
($\sigma^2 = 0.5$)

	$\lambda_1 = 0.1$	$\lambda_1 = 0.5$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
ML				
T=30	3872516.900	6778576.000	19123439.000	3388751.100
T=60	694515.670	300249.240	469677.400	428907.920
RML				
T=30	15.681	534.208	1701.617	4274.412
T=60	34.074	1239.192	3908.305	9728.462
Pretest^1				
T=30	1899908.200	6439488.700	19014454.000	3356635.800
T=60	513906.270	299499.740	469757.230	429015.270
Pretest^2				
T=30	3350379.700	6731594.200	19106842.000	3386574.800
T=60	669411.800	300068.230	469727.670	428987.160
Pretest^3				
T=30	3644051.400	6760342.100	19116509.000	3388537.900
T=60	678897.450	300145.440	469694.110	428942.990
Stein^1				
T=30	104154.210	2045560.700	4528802.600	1072963.800
T=60	71568.871	148850.690	272856.030	226707.880
Stein^2				
T=30	405135.890	3072453.900	7907491.300	1620836.400
T=60	150494.240	186783.180	326102.440	281732.790
Stein^3				
T=30	700173.920	3645767.700	9733764.800	1912433.500
T=60	215306.820	206164.060	351927.750	308309.540

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.5: The Risk Ratios Relative to the ML Estimator
— Quadratic Loss: $\sigma^2 = 0.1$ —

	$\lambda_1 = 0.1$	$\lambda_1 = 0.5$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
RML				
T=30	0.157	0.537	0.612	0.504
T=60	0.645	2.109	2.080	1.866
Pretest^1				
T=30	0.897	1.003	1.000	1.000
T=60	1.054	1.000	1.000	1.000
Pretest^2				
T=30	0.980	1.001	1.000	1.000
T=60	1.025	1.000	1.000	1.000
Pretest^3				
T=30	0.990	1.000	1.000	1.000
T=60	1.012	1.000	1.000	1.000
Stein^1				
T=30	0.264	0.531	0.581	0.590
T=60	0.533	0.778	0.776	0.707
Stein^2				
T=30	0.378	0.632	0.675	0.691
T=60	0.570	0.808	0.821	0.776
Stein^3				
T=30	0.453	0.690	0.727	0.743
T=60	0.608	0.833	0.847	0.813

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.6: The Risk Ratios Relative to the ML Estimator
— Quadratic Loss: $\sigma^2 = 0.5$ —

	$\lambda_1 = 0.1$	$\lambda_1 = 0.5$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
RML				
T=30	0.001	0.002	0.001	0.005
T=60	0.009	0.066	0.058	0.098
Pretest^1				
T=30	0.454	0.941	0.993	0.993
T=60	0.710	1.001	1.002	1.002
Pretest^2				
T=30	0.852	0.991	0.999	1.000
T=60	0.943	1.001	1.001	1.001
Pretest^3				
T=30	0.925	0.996	1.000	1.000
T=60	0.966	1.000	1.000	1.000
Stein^1				
T=30	0.022	0.265	0.227	0.284
T=60	0.099	0.497	0.574	0.535
Stein^2				
T=30	0.093	0.418	0.404	0.451
T=60	0.208	0.616	0.686	0.658
Stein^3				
T=30	0.166	0.506	0.501	0.541
T=60	0.298	0.679	0.741	0.718

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.7: The Risk Ratios Relative to the ML Estimator
— Weighted Quadratic Loss: $\sigma^2 = 0.1$ —

	$\lambda_1 = 0.1$	$\lambda_1 = 0.5$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
RML				
T=30	0.001	0.039	0.107	0.217
T=60	0.017	0.662	1.746	3.340
Pretest^1				
T=30	0.889	1.000	1.000	1.000
T=60	0.966	1.000	1.000	1.000
Pretest^2				
T=30	0.978	1.000	1.000	1.000
T=60	0.992	1.000	1.000	1.000
Pretest^3				
T=30	0.991	1.000	1.000	1.000
T=60	0.996	1.000	1.000	1.000
Stein^1				
T=30	0.186	0.467	0.542	0.620
T=60	0.264	0.688	0.816	0.874
Stein^2				
T=30	0.341	0.608	0.664	0.722
T=60	0.420	0.766	0.852	0.884
Stein^3				
T=30	0.434	0.678	0.724	0.771
T=60	0.508	0.806	0.874	0.896

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.8: The Risk Ratios Relative to the ML Estimator
— Weighted Quadratic Loss: $\sigma^2 = 0.5$ —

	$\lambda_1 = 0.1$	$\lambda_1 = 0.5$	$\lambda_1 = 1.0$	$\lambda_1 = 2.0$
RML				
T=30	0.000	0.000	0.000	0.001
T=60	0.000	0.004	0.008	0.023
Pretest^1				
T=30	0.491	0.950	0.994	0.991
T=60	0.740	0.998	1.000	1.000
Pretest^2				
T=30	0.865	0.993	0.999	0.999
T=60	0.964	0.999	1.000	1.000
Pretest^3				
T=30	0.941	0.997	1.000	1.000
T=60	0.978	1.000	1.000	1.000
Stein^1				
T=30	0.027	0.302	0.237	0.317
T=60	0.103	0.496	0.581	0.529
Stein^2				
T=30	0.105	0.453	0.413	0.478
T=60	0.217	0.622	0.694	0.657
Stein^3				
T=30	0.181	0.538	0.509	0.564
T=60	0.310	0.687	0.749	0.719

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.9: Estimated MSE for Parameter Estimators
($\sigma^2 = 0.1$, $\lambda_1 = 0.1$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
ML						
T=30	29.666	39.198	0.805	0.003	0.273	0.069
T=60	12.040	5.803	0.166	0.002	0.141	0.028
RML						
T=30	10.776	0.130	0.039	0.010	0.010	0.000
T=60	11.542	0.126	0.043	0.010	0.010	0.000
Pretest^1						
T=30	27.825	34.060	0.699	0.007	0.184	0.045
T=60	13.394	5.453	0.163	0.004	0.121	0.023
Pretest^2						
T=30	29.654	37.909	0.772	0.005	0.230	0.058
T=60	12.592	5.712	0.166	0.002	0.134	0.027
Pretest^3						
T=30	29.728	38.457	0.792	0.004	0.249	0.062
T=60	12.303	5.755	0.167	0.002	0.138	0.027
Stein^1						
T=30	11.802	6.503	0.154	0.007	0.033	0.006
T=60	8.279	1.319	0.050	0.005	0.029	0.005
Stein^2						
T=30	13.705	12.454	0.257	0.006	0.062	0.014
T=60	8.049	2.177	0.068	0.004	0.047	0.009
Stein^3						
T=30	15.182	16.097	0.324	0.005	0.084	0.020
T=60	8.197	2.697	0.080	0.003	0.059	0.011

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.10: Estimated MSE for Parameter Estimators
($\sigma^2 = 0.1$, $\lambda_1 = 0.5$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
ML						
T=30	42.344	37.178	1.225	0.017	0.274	0.063
T=60	15.740	4.915	0.223	0.010	0.139	0.028
RML						
T=30	42.065	1.070	0.145	0.250	0.010	0.000
T=60	42.904	1.090	0.150	0.250	0.010	0.000
Pretest^1						
T=30	42.565	37.184	1.226	0.019	0.274	0.063
T=60	15.740	4.915	0.223	0.010	0.139	0.028
Pretest^2						
T=30	42.385	37.179	1.225	0.018	0.274	0.063
T=60	15.740	4.915	0.223	0.010	0.139	0.028
Pretest^3						
T=30	42.344	37.178	1.225	0.017	0.274	0.063
T=60	15.740	4.915	0.223	0.010	0.139	0.028
Stein^1						
T=30	26.392	15.902	0.556	0.056	0.104	0.024
T=60	13.330	2.789	0.136	0.028	0.078	0.016
Stein^2						
T=30	28.870	21.446	0.725	0.035	0.146	0.034
T=60	13.368	3.361	0.158	0.019	0.095	0.019
Stein^3						
T=30	30.673	24.206	0.811	0.028	0.168	0.039
T=60	13.590	3.640	0.169	0.016	0.103	0.021

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.11: Estimated MSE for Parameter Estimators
($\sigma^2 = 0.1$, $\lambda_1 = 1.0$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
ML						
T=30	58.005	39.829	1.673	0.045	0.278	0.055
T=60	23.434	5.746	0.322	0.031	0.139	0.027
RML						
T=30	58.419	1.512	0.186	1.000	0.010	0.000
T=60	59.040	1.531	0.190	1.000	0.010	0.000
Pretest^1						
T=30	58.005	39.829	1.673	0.045	0.278	0.055
T=60	23.434	5.746	0.322	0.031	0.139	0.027
Pretest^2						
T=30	58.005	39.829	1.673	0.045	0.278	0.055
T=60	23.434	5.746	0.322	0.031	0.139	0.027
Pretest^3						
T=30	58.005	39.829	1.673	0.045	0.278	0.055
T=60	23.434	5.746	0.322	0.031	0.139	0.027
Stein^1						
T=30	36.446	20.492	0.789	0.164	0.123	0.024
T=60	19.034	3.613	0.214	0.076	0.090	0.017
Stein^2						
T=30	40.419	25.688	1.019	0.099	0.163	0.032
T=60	19.746	4.204	0.243	0.053	0.104	0.020
Stein^3						
T=30	42.964	28.222	1.134	0.078	0.183	0.036
T=60	20.250	4.486	0.257	0.045	0.110	0.021

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.12: Estimated MSE for Parameter Estimators
($\sigma^2 = 0.1$, $\lambda_1 = 2.0$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
ML						
T=30	85.239	67.436	2.299	0.141	0.271	0.052
T=60	35.165	6.273	0.465	0.092	0.147	0.028
RML						
T=30	72.274	1.829	0.214	4.000	0.010	0.000
T=60	72.634	1.844	0.217	4.000	0.010	0.000
Pretest^1						
T=30	85.239	67.436	2.299	0.141	0.271	0.052
T=60	35.165	6.273	0.465	0.092	0.147	0.028
Pretest^2						
T=30	85.239	67.436	2.299	0.141	0.271	0.052
T=60	35.165	6.273	0.465	0.092	0.147	0.028
Pretest^3						
T=30	85.239	67.436	2.299	0.141	0.271	0.052
T=60	35.165	6.273	0.465	0.092	0.147	0.028
Stein^1						
T=30	52.149	37.604	1.290	0.470	0.137	0.026
T=60	24.951	4.250	0.305	0.205	0.099	0.018
Stein^2						
T=30	59.488	45.840	1.563	0.278	0.173	0.033
T=60	27.300	4.820	0.349	0.139	0.113	0.021
Stein^3						
T=30	63.584	49.783	1.696	0.219	0.190	0.037
T=60	28.578	5.088	0.370	0.119	0.119	0.022

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.13: Estimated MSE for Parameter Estimators
($\sigma^2 = 0.5$, $\lambda_1 = 0.1$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
ML						
T=30	1548.344	10226.516	47.122	0.008	1.408	0.276
T=60	253.792	1085.586	3.837	0.004	0.856	0.099
RML						
T=30	10.833	0.172	0.039	0.010	0.010	0.000
T=60	11.731	0.145	0.043	0.010	0.010	0.000
Pretest^1						
T=30	662.768	4674.333	28.440	0.011	0.304	0.074
T=60	172.813	777.610	3.398	0.009	0.277	0.033
Pretest^2						
T=30	1262.803	8777.437	38.379	0.010	0.644	0.140
T=60	231.841	1031.629	3.692	0.008	0.559	0.060
Pretest^3						
T=30	1396.535	9498.818	45.368	0.010	0.818	0.180
T=60	240.999	1052.773	3.748	0.007	0.656	0.071
Stein^1						
T=30	48.352	207.207	3.263	0.009	0.032	0.008
T=60	33.278	98.359	1.047	0.009	0.033	0.004
Stein^2						
T=30	154.700	930.993	8.505	0.008	0.085	0.022
T=60	59.621	217.538	1.594	0.008	0.084	0.010
Stein^3						
T=30	264.256	1682.476	12.535	0.008	0.143	0.036
T=60	81.046	317.869	1.909	0.007	0.135	0.016

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.14: Estimated MSE for Parameter Estimators
($\sigma^2 = 0.5$, $\lambda_1 = 0.5$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
ML						
T=30	2546.285	17133.708	105.546	0.045	1.413	0.254
T=60	160.387	507.897	3.920	0.024	0.785	0.099
RML						
T=30	42.170	1.071	0.144	0.250	0.010	0.000
T=60	43.042	1.084	0.149	0.250	0.010	0.000
Pretest^1						
T=30	2395.085	16118.141	103.814	0.134	1.003	0.158
T=60	162.862	505.856	3.900	0.059	0.698	0.086
Pretest^2						
T=30	2519.928	16988.129	104.682	0.088	1.212	0.210
T=60	1612.212	507.453	3.912	0.036	0.763	0.095
Pretest^3						
T=30	2534.058	17072.343	105.445	0.072	1.289	0.229
T=60	160.683	507.526	3.922	0.031	0.769	0.096
Stein^1						
T=30	738.791	4452.678	48.049	0.171	0.169	0.025
T=60	84.822	247.681	1.567	0.113	0.188	0.023
Stein^2						
T=30	1120.189	7086.298	62.654	0.127	0.336	0.053
T=60	100.842	311.214	2.133	0.075	0.305	0.037
Stein^3						
T=30	1336.407	8599.740	69.968	0.106	0.453	0.074
T=60	109.805	344.239	2.428	0.060	0.374	0.046

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.15: Estimated MSE for Parameter Estimators
($\sigma^2 = 0.5$, $\lambda_1 = 1.0$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
ML						
T=30	7342.337	52830.039	323.787	0.123	1.358	0.250
T=60	253.103	807.029	5.012	0.076	0.832	0.108
RML						
T=30	58.371	1.512	0.185	1.000	0.010	0.000
T=60	59.123	1.524	0.189	1.000	0.010	0.000
Pretest^1						
T=30	7279.439	52483.712	322.622	0.394	1.006	0.201
T=60	255.335	806.649	5.017	0.136	0.806	0.104
Pretest^2						
T=30	7330.033	52771.631	323.674	0.233	1.225	0.237
T=60	253.722	807.039	5.014	0.092	0.826	0.107
Pretest^3						
T=30	7337.470	52805.612	323.792	0.179	1.278	0.243
T=60	253.249	807.034	5.013	0.080	0.830	0.108
Stein^1						
T=30	1784.831	11869.734	92.246	0.577	0.194	0.040
T=60	144.998	464.568	2.193	0.330	0.269	0.035
Stein^2						
T=30	3060.176	21253.822	147.969	0.400	0.365	0.075
T=60	170.962	556.655	2.910	0.212	0.395	0.052
Stein^3						
T=30	3752.553	26362.272	177.262	0.317	0.482	0.098
T=60	184.528	601.518	3.272	0.167	0.464	0.061

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

Table 5.16: Estimated MSE for Parameter Estimators
($\sigma^2 = 0.5$, $\lambda_1 = 2.0$)

	β_1	β_2	β_3	λ_1	λ_2	λ_3
ML						
T=30	2042.220	14538.089	228.632	0.331	1.390	0.200
T=60	270.991	517.690	11.944	0.224	0.823	0.101
RML						
T=30	72.312	1.823	0.214	4.000	0.010	0.000
T=60	72.634	1.840	0.216	4.000	0.010	0.000
Pretest^1						
T=30	2029.834	14438.281	227.593	1.125	1.188	0.168
T=60	272.545	517.434	11.935	0.366	0.808	0.098
Pretest^2						
T=30	2042.449	14529.497	228.622	0.552	1.337	0.193
T=60	271.452	517.696	11.945	0.258	0.821	0.100
Pretest^3						
T=30	2043.062	14536.437	228.627	0.441	1.365	0.197
T=60	271.122	517.693	11.944	0.236	0.823	0.101
Stein^1						
T=30	676.338	4036.364	54.591	2.065	0.250	0.035
T=60	150.905	271.014	6.009	1.013	0.327	0.039
Stein^2						
T=30	994.254	6488.944	95.121	1.337	0.449	0.064
T=60	180.646	337.783	7.605	0.634	0.445	0.054
Stein^3						
T=30	1164.479	7807.121	116.934	1.017	0.576	0.082
T=60	195.913	370.171	8.382	0.498	0.507	0.061

RML = constrained ML estimates
Pretest^1 = pretest estimates at level 0.01
Pretest^2 = pretest estimates at level 0.05
Pretest^3 = pretest estimates at level 0.10
Stein^1 = shrinkage estimates with $a = \chi^2_3(0.01)$
Stein^2 = shrinkage estimates with $a = \chi^2_3(0.05)$
Stein^3 = shrinkage estimates with $a = \chi^2_3(0.10)$

CHAPTER 6

CONCLUDING REMARKS

April is the cruellest month, breeding
Lilacs out of the dead land, mixing
Memory and desire, stirring
Dull roots with spring rain.
Winter kept us warm, covering
Earth in forgetful snow, feeding
A little life with dried tubers.
Summer surprised us, coming over the Starnbergersee
With a shower of rain; we stopped in the colonnade,
And went on in the sunlight, into the Hofgarten,
And drank coffee, and talked for an hour.
Bin gar keine Russin, stamm' aus Litauen, echt deutsch
And when we were children, staying at the arch-duke's,
My cousin's, he took me out on a sled,
And I was frightened. He said, Marie,
Marie, hold on tight. And down we went.
In the mountains, there you feel free.
I read, much of the night, and go south in the winter.

T. S. Eliot

In this dissertation estimation and inference procedures for the Box-Cox regression model have been considered. This model is an important one for economic researchers, as it allows the joint estimation of response coefficients and parameters that determine the functional form. Since the model is nonlinear in the parameters, maximum likelihood estimation is employed, and the focus of this research has been the sampling characteristics of estimators and test statistics in small samples.

Zarembka (1974) has shown that heteroscedasticity induced by the transformation leads to bias in the maximum likelihood estimator of the power parameter on the response variable. However, the normality assumption for the error disturbances is not valid in a strict sense, since the Box-Cox transformation requires a

positive value for the pretransformed variable, which is thereby truncated. Since the heteroscedasticity analyzed by Zarembka is based on the normality of the error term, his condition for consistent estimation of the power parameter of the response variable is not, in general, correct. This research addressed the truncation problem since the validity of ML estimation and hypothesis testing is based on the assumption of a regular likelihood function.

Chapter 2 dealt with the basic concepts of statistical inference with an emphasis on ML estimation and asymptotic properties of the ML estimators. Finite sample properties and the linear regression model were briefly mentioned. Biased estimators (pretest and Stein-rule) were given a rather thorough treatment within the framework of decision theory. The small sample distributions of estimators and test statistics were discussed in light of current theoretical developments.

In Chapter 3, the general transformation-of-variables model was presented. The three methods of estimation — ML, two stage nonlinear least squares and IGLS — were presented. Bootstrapping was carried out to compare the bootstrap sample variability and the finite sample variability (RMSE). When $T = 30$, bootstrap sample variabilities turned out to be poor approximations to finite sample variabilities. The exact distribution of the t statistic was obtained using an Edgeworth expansion. To correct the size of the t test, bootstrap inversion of an Edgeworth expansion was employed. Our Monte Carlo experiments showed that the size correction using the Edgeworth expansion, which was approximated

by bootstrapping, improved the performance of the t-test. The biases of linear (response) parameter estimators, especially those with nonlinear power transformations, were shown to be substantial when the sample size is small ($T = 30$). The standard errors computed from the Hessian matrix were a poor measure of the finite sample variability. The t-ratios of the linear parameter estimators may not be normally distributed in small samples.

The LM, LR and Wald test statistics can be used for testing the functional form in the Box-Cox model. Though they all have an asymptotic χ^2-distribution, the empirical size of the tests does not correspond to the nominal size. Furthermore, these statistics do not have a χ^2-distribution in small samples. Chapter 4 is primarily concerned with hypothesis testing within the framework of the Box-Cox transformation. In order to insure the accuracy of our analysis, we assumed that the distribution of the pretransformed response variable was truncated normal. The asymptotic equality of the LM, LR and Wald statistics was displayed. The Monte Carlo simulation, however, yielded differences among the three statistics when $T = 30$ and $T = 60$. The LR statistic appeared to be a good approximation to the χ^2-distribution in small samples. The asymptotic power functions of the LM, LR and Wald statistics were shown to be a poor measure of the small sample power.

In Chapter 5, the asymptotic risk properties of the ML, constrained ML, pretest and shrinkage estimators were analyzed. The shrinkage (positive-part Stein-like)

estimator dominated the ML estimator in small samples as well as in large samples.

Basically, our research is based on the assumptions of independent random disturbances and nonstochastic regressors while often in practice the explanatory variables contain lagged dependent variables and the error term is autocorrelated. In addition, the available test statistics are only asymptotically justified. Therefore, future research topics are suggested as follows:

1. When we depend on asymptotic test statistics, it is desirable to develop size corrected test procedures using an Edgeworth expansion. Bootstrapping is an alternative way of correcting the size of the asymptotic test statistics.

2. It is required to obtain a measure of the precision of the shrinkage estimator in the nonlinear regression model. In addition, the finite risk needs to be investigated since the risk function has been obtained only asymptotically.

3. When the Box-Cox regression model contains lagged dependent variables with autocorrelated errors, the usual asymptotic covariance matrix of parameter estimators is incorrect. In order to analyze dynamic nonlinear models, the notions of asymptotic martingales, which are called mixingales, and near epoch dependence need to be introduced [Gallant (1987); Gallant and White (1988)], since these concepts enable us to investigate the general nonlinear model with dependent and heterogenous stochastic processes. Furthermore, Gallant and White's unified approach to nonlinear dynamic models provides

a theoretical foundation for acceptable statistical inference even in the presence of model misspecification.

4. Consider the general transformation-of-variables model:

$$T(y_t) = \beta_1 + \beta_2 x_{2t}^{(\lambda_2)} + , \ldots , + \beta_k x_{kt}^{(\lambda_k)} + \epsilon_t$$

where $T(y_t)$ is a suitable transformation that makes the random variables y_t normally distributed and $x_{it}, i = 1, \ldots, k$, are assumed to be nonstochastic and $E\epsilon_t = 0$. Using Bartlett's (1947) method for the approximation [Box and Hill (1974)], we obtain

$$\text{Var}[T(y_t)] = \text{Var}(y_t)[\tfrac{\partial T(y_t)}{\partial y_t}]^2_{y_t = Ey_t}$$

Therefore, $\text{Var}[T(y_t)]$ might be heteroscedastic even if $\text{Var}(y_t)$ is homoscedastic. For the Box-Cox transformation $y_t^{(\lambda_1)}$ for $L < T(y_t) < R$,

$$\text{Var}[y_t^{(\lambda_1)}] = \text{Var}(y_t)[E(y_t)]^{2\lambda_1 - 2} = \text{Var}(\epsilon_t)$$

where $R = -\frac{1}{\lambda_1}$ and $L = -\infty$ if $\lambda_1 < 0$; $R = +\infty$ and $L = -\frac{1}{\lambda_1}$ if $\lambda_1 > 0$. The truncated ML estimator under the assumption of homoscedasticity is inconsistent when the true error disturbances have heteroscedastic variances [Hurd (1979)]. In general, heteroscedasticity caused by the transformation needs to be considered for efficient estimation, even though the ML estimator is consistent.

5. In the context of the Box-Cox transformation, the sample can be divided into two or more subsamples where all or some of parameters (linear and

power) are different. For example, in cross-sectional analysis, the regression coefficients and functional form, as well as variances of the error term, might differ in geographical regions. Then the problem is that the number of observations might be small relative to the number of parameters estimated. A way to cope with this difficulty is to apply the Lagrange principle to test for structural change, since it requires only the estimation of the constrained model. If the sample size is very small, the size correction can be made using an Edgeworth expansion.

6. Spitzer (1977) estimated simultaneous equations for money demand and supply, when both the money demand and supply equations are transformed. He used the full information ML (FIML) estimation method. The truncation problem raised in single equation estimation can also be addressed in the simultaneous equations framework. In this case, the covariance matrices computed from the Hessian matrix or outer products of the first derivatives of the usual log-likelihood function are incorrect. There are two more points worthy of mentioning. First, Spitzer ignored the identification conditions in nonlinear simultaneous equation systems by relying on the identification conditions for equations linear in the parameters, but nonlinear in the data. Second, he estimated the simultaneous Box-Cox model using a small sample ($T = 41$), for which the validity of applying asymptotic properties is doubtful. Spitzer's simultaneous Box-Cox model could be extended when the error

disturbances are assumed to be heteroscedastic and/or vector autoregressive.

7. Carroll and Ruppert (1984) employed the Box-Cox transformation when fitting a theoretical model to data. If we have the following CES production function:

$$y = [\delta_0 + \delta_1 K^{\rho} + \delta_2 L^{\rho}]^{1/\rho}$$

economists usually model the data generation process by merely adding an error term:

$$y_t = [\delta_0 + \delta_1 K_t^{\rho} + \delta_2 L_t^{\rho}]^{1/\rho} + \epsilon_t \tag{6.1}$$

or

$$y_t^{\rho} = \delta_0 + \delta_1 K_t^{\rho} + \delta_2 L_t^{\rho} + \epsilon_t \tag{6.2}$$

where $\epsilon_t \sim N(0, \sigma^2)$. The model (6.1) is estimated by nonlinear least squares or the ML method while model (6.2) is estimated by the ML method. Carroll and Ruppert have proposed that the simultaneous power transformation of the response variable and the model lead to homoscedasticity and normality of error disturbances:

$$y_t^{(\lambda)} = \{[\delta_0 + \delta_1 K^{\rho} + \delta_2 L^{\rho}]^{1/\rho}\}^{(\lambda)} + \epsilon_t, \quad \epsilon_t \sim \text{iid } N(0, \sigma^2) \tag{6.3}$$

They further showed via Monte Carlo simulation and an asymptotic theory that the cost of ML estimation of λ is moderate compared to the case when λ is known. Therefore, the method of transformation suggested by Carroll and Ruppert merits attention because we can apply ML estimation to estimate

the random data generating mechanism without destroying the theoretical relationship. The Box-Cox transformation of the truncated dependent variable leads to a truncated transformed dependent variable. Therefore, under the assumption of an appropriate distribution function for the original dependent variable, we can get accurate ML estimators. Heteroscedasticity caused by transformation can also be implemented into model (6.3).

8. Since the error disturbances of the Box-Cox transformation model are assumed to be approximately normal, robust estimation methods can be used when we expect heavier tail behaviour of the error disturbances. Robust estimators for the Box-Cox model were proposed by Carroll (1980) and Bickel and Doksum (1981) when only the dependent variable is transformed. Further study is required to extend robust estimation to the general transformation-of-variables model. Han (1987) proposed a non-parametric estimation method that attains efficiency gain in terms of mean squared error over the ML estimation method. Neither the asymptotic nor the finite properties of the robust estimators have been studied for hypothesis testing.

Bibliography

[1] Adkins, L. C. and Hill, R. C. (1989). "Risk Characteristics of a Stein-Like Estimator for the Probit Regression Model." *Economics Letters*, 30, 19–26.

[2] Aitchison, J. and Silvey, S. D. (1958). "Maximum-Likelihood Estimation of Parameters Subject to Restraints." *Annals of Mathematical Statistics*, 29, 813–828.

[3] Amemiya, T. (1985). *Advanced Econometrics.* Oxford: Basil Blackwell.

[4] Amemiya, T. and Powell, J. L. (1981). "A Comparison of the Box-Cox Maximum Likelihood Estimator and the Nonlinear Two Stage Least Squares Estimator." *Journal of Econometrics*, 17, 351–381.

[5] Aneuryn-Evans, G. and Deaton, A. (1980). "Testing Linear versus Logarithmic Regression Models." *Review of Economic Studies*, 57, 275–291.

[6] Angus, J. E. (1989). "A Note on the Central Limit Theorem for the Bootstrap Mean." *Communications in Statistics — Theory and Methods*, 18, 1979–1982.

[7] Baranchik, A. J. (1970). "A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution." *Annals of Mathematical Statistics*, 41, 642–645.

[8] Barnett, W. A. (1976). "Maximum Likelihood and Iterative Aitken Estimation of Nonlinear Systems of Equations." *Journal of the American Statistical Association*, 71, 354–360.

[9] Bartlett, M. S. (1937). "Properties of Sufficiency and Statistical Tests." *Proceedings of the Royal Society of London*, ser A, 160, 268–282.

[10] Bartlett, M. S. (1947). "The Use of Transformations." *Biometrics*, 3, 39–52.

[11] Basman, K. L. (1972). "Exact Finite Sample Distributions for Some Econometric Estimators and Test Statistics: A Survey and Appraisal," in *Frontiers in Quantitative Economics.* Edited by M. D. Intriligator and D. A. Kendrick. Amsterdam: North-Holland.

[12] Beran, R. J. (1982). "Estimated Sampling Distributions: Bootstrap and Competitors." *Annals of Statistics*, 10, 212-225.

[13] Bergstrom, A. R. (1962). "The Exact Sampling Distributions of Least Squares and Maximum Likelihood Estimators of the Marginal Propensity to Consume." *Econometrica*, 30, 480–490.

[14] Berndt, E. K., Hall, B. H., Hall, R. E., and Hausman, J. A. (1974). "Estimation and Inference in Nonlinear Structural Models." *Annals of Economic and Social Measurement*, 3, 653–665.

[15] Bickel, P. J. and Doksom, K. A. (1981). "An Analysis of Transformations Revisited." *Journal of the American Statistical Association*, 76, 296–311.

[16] Bickel, P. J. and Freedman, D. A. (1981). "Some Asymptotic Theory for the Bootstrap." *Annals of Statistics*, 9, 1196–1217.

[17] Birens, H. (1982). "Consistent Model Specification Tests." *Journal of Econometrics*, 26, 323–353.

[18] Blackley, P., Follain, J. R., Jr., and Ondrich, J. (1984). "Box-Cox Estimation of Hedonic Models: How Serious Is the Iterative OLS Variance Bias ?." *The Review of Economics and Statistics*, 66, 348–352.

[19] Blaylock, J. R. and Smallwood, D. M. (1985). "Box-Cox Transformation and a Heteroskedastic Error Variance: Import Demand Equation Revisited." *International Statistical Review*, 53, 91–97.

[20] Box, G. E. P. and Cox, D. R. (1964). "An Analysis of Transformation." *Journal of Royal Statistical Society*, ser B, 26, 211–243.

[21] Box, G. E. P. and Cox, D. R. (1982). "An Analysis of Transformations Revisited, Rebutted." *Journal of the American Statistical Association*, 77, 209–210.

[22] Box, G. E. P. and Hill, W. J. (1974). "Correcting Inhomogeneity of Variance with Power Transformation Weighting." *Technometrics*, 16, 385–389.

[23] Box, G. E. P. and Tidwell, P. W. (1962). "Transformation of the Independent Variables." *Technometrics*, 4, 531–550.

[24] Boylan, T. A., Cuddy, M. P., and O'Muircheartaigh, I. G. (1982). "Import Demand Equations: An Application of a Generalized Box-Cox Methodology." *International Statistical Review*, 50, 103–112.

[25] Boylan, T. A. and O'Muircheartaigh, I. G. (1981). "The Functional Form of the U.K. Demand for Money: A Critique of a Paper by Mills." *Applied Statistics*, 30, 296–299.

[26] Brandwein, A. C. and Strawderman, W. E. (1978). "Minimax Estimation of Location Parameters for Spherically Symmetric Unimodal Distributions under Quadratic Loss." *Annals of Statistics*, 6, 377–416.

[27] Brandwein, A. C. and Strawderman, W. E. (1980). "Minimax Estimation of Location Parameters for Spherically Symmetric Distributions with Concave Loss." *Annals of Statistics*, 8, 279–284.

[28] Breusch, T. S. and Pagan, A. R. (1980). "The Lagrangian Multiplier Test and Its Applications to Model Specification in Econometrics." *Review of Economic Studies*, 57, 239–253.

[29] Calzolari, R. J. and Panattoni, L. (1984). "Alternative Estimators of FIML Covariance Matrix: A Monte Carlo Study." *Econometrica*, 56, 701–714.

[30] Carroll, R. J. and Ruppert, D. (1984). "Power Transformations When Fitting Theoretical Models to Data." *Journal of the American Statistical Association*, 79, 321–328.

[31] Copas, J. B. (1983). "Regression, Prediction and Shrinkage." *Journal of Royal Statistical Society*, ser B, 45, 311–354.

[32] Cox, D. R. (1984). "Effective Degrees of Freedom and Likelihood the Likelihood Ratio Tests." *Biometrika*, 71, 487–493.

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

[34] Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

[35] Cramer, J. S. (1986). *Econometric Applications of Maximum Likelihood Method*. Cambridge: Cambridge University Press.

[36] Dagenais, M. G. (1978). "The Computation of FIML Estimates in Linear and Nonlinear Simultaneous Equations Models." *Econometrica*, 46, 1351–1362.

[37] Dagenais, M. G. (1983). "Extension of the Ridge Regression Technique to Non-Linear Models with Additive Errors." *Economics Letters*, 12, 169–174.

[38] Daniels, H. E. (1954). "Saddlepoint Approximations in Statistics." *Annals of Mathematical Statistics*, 25, 631–650.

[39] Daniels, H. E. (1980). "Exact Saddlepoint Approximations." *Biometrika*, 67, 59–63.

[40] Davidson, R. and MacKinnon, J. G. (1983). "Small Sample Properties of Alternative Forms of the Lagrange Multiplier Test." *Economics Letters*, 12, 269–275.

[41] Davidson, R. and MacKinnon, J. G. (1985). "Testing Linear and Loglinear Regressions against Box-Cox Alternatives." *Canadian Journal of Economics*, 18, 499–517.

[42] Dennis, J. E., Jr. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Englewood Cliffs, New Jersey: Prentice-Hall.

[43] Draper, N. R. and Cox, D. R. (1969). "On Distributions and Their Transformation to Normality." *Journal of Royal Statistical Society*, ser B, 31, 472–476.

[44] Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics*, 7, 1–26.

[45] Efron, B. (1982a). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia: S.I.A.M.

[46] Efron, B. (1982b). "Maximum Likelihood and Decision Theory." *Annals of Statistics*, 10, 340–356.

[47] Efron, B. and Gong, G. (1983). "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation." *The American Statistician*, 37, 36–48.

[48] Efron, B. and Hinkley, D. V. (1978). "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed versus Expected Fisher Information." *Biometrika*, 65, 457–487.

[49] Efron, B. and Morris, C. (1973). "Stein's Estimation Rule and Its Competitors — An Empirical Bayes Approach." *Journal of the American Statistical Association*, 68, 117–130.

[50] Efron, B. and Morris, C. (1975). "Data Analysis Using Stein's Estimator and Its Generalizations." *Journal of the American Statistical Association*, 70, 311–319.

[51] Efron, B. and Morris, C. (1976). "Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution." *Annals of Statistics*, 4, 11–21.

[52] Engle, R. F. (1984). "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," in *Handbook of Econometrics.* Edited by Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

[53] Evans, G. B. A. and Savin, N. E. (1982). "Conflict among the Criteria Revisited: The W, LR and LM Tests." *Econometrica*, 50, 737–748.

[54] Fletcher, R. and Powell, M. J. D. (1963). "A Rapidly Convergent Descent Method for Minimization." *The Computer Journal*, 6, 163–168.

[55] Fomby, T. B., Hill, R. C., and Johnson, S. R. (1985). *Advanced Econometric Methods.* New York: Springer-Verlag.

[56] Freedman, D. A. (1981). "Bootstrapping Regression Models." *Annals of Statistics*, 9, 1218–1228.

[57] Freedman, D. A. (1984). "On Bootstrapping Two-Stage Least-Squares Estimates in Stationary Linear Models." *Annals of Statistics*, 12, 827–842.

[58] Freedman, D. A. and Peters, S. C. (1984a). "Bootstrapping a Regression Equation: Some Empirical Results." *Journal of the American Statistical Association*, 79, 97–106.

[59] Freedman, D. A. and Peters, S. C. (1984b). "Bootstrapping an Econometric Model: Some Empirical Results." *Journal of Business and Economic Statistics*, 2, 150–158.

[60] Gallant, A. R. (1987). *Nonlinear Statistical Models.* New York: John Wiley & Sons.

[61] Gallant, A. R. and White, H. (1988). *A Unified Theory of Estimation & Inference for Nonlinear Dynamic Models.* Oxford: Basil Blackwell.

[62] Gill, P. E., Murray, W., and Wright, M. H. (1981). *Practical Optimization.* New York: Academic Press.

[63] Godfrey, L. G. (1978a). "Testing Against General Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables." *Econometrica*, 46, 1293–1302.

[64] Godfrey, L. G. (1978b). "Testing for Higher Order Serial Correlation in Regression Equations When the Regressors Include Lagged Dependent Variables." *Econometrica*, 46, 1303–1310.

[65] Godfrey, L. G. (1981). "On the Invariance of the Lagrange Multiplier Test with respect to Certain Changes in the Alternative Hypothesis." *Econometrica*, 49, 1443–1455.

[66] Godfrey, L. G. (1988). *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches.* Cambridge: Cambridge University Press.

[67] Goldfeld, S. M. and Quandt, R. E. (1965). *Nonlinear Methods in Econometrics.* Amsterdam: North-Holland.

[68] Granger, C. W. J. and Newbold, P. (1976). "Forecasting Transformed Series." *Journal of Royal Statistical Society*, ser B, 38, 189–203.

[69] Gregory, A. W. and Veall, M. R. (1985). "Formulating Wald Tests of Nonlinear Restrictions." *Econometrica*, 53, 1465–1468.

[70] Griffiths, W. E., Hill, R. C., and Pope, P. J. (1987). "Small Sample Properties of Probit Model Estimators." *Journal of the American Statistical Association*, 82, 929–937.

[71] Guerrero, V. M. (1987). "A Note on the Estimation of Atkinson's Index of Inequality." *Economics Letters*, 25, 379–384.

[72] Hall, P. (1983). "Inverting an Edgeworth Expansion." *Annals of Statistics*, 11, 569–576.

[73] Hauck, W. W. and Donner, A. (1977). "Wald's Test as Applied to Hypothesis in Logit Analysis." *Journal of the American Statistical Association*, 72, 851–853.

[74] Hendry, D. F. (1984). "Monte Carlo Experimentation in Econometrics," in *Handbook of Econometrics*. Edited by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland.

[75] Hernandez, F. and Johnson, R. A. (1980). "The Large-Sample Behavior of Transformations to Normality." *Journal of the American Statistical Association*, 75, 855–861.

[76] Hinkley, D. V. (1975). "On Power Transformations to Symmetry." *Biometrika*, 62, 101–111.

[77] Hinkley, D. V. (1988). "Bootstrap Methods." *Journal of Royal Statistical Society*, ser B, 50, 321–337.

[78] Hinkley, D. V. and Runger, G. (1984). "The Analysis of Transformed Data." *Journal of the American Statistical Association*, 79, 302–309.

[79] Hogg, R. V. and Craig, A. T. (1978). *Introduction to Mathematical Statistics*. 4th ed. New York: MacMillan.

[80] Holly, A. and Phillips, P. C. B. (1979). "A Saddlepoint Approximation to the Distribution of the k-Class Estimator of a Coefficient in a Simultaneous System." *Econometrica*, 47, 1527–1547.

[81] Hopwood, W. S., McKeown, J. C., and Newbold, P. (1984). "Time Series Forecasting Models Involving Power Transformations." *Journal of Forecasting*, 3, 57–61.

[82] Hwang, H.-S. (1981). "Demand for Money: Tests of Functional Forms and Stability." *Applied Economics*, 13, 235–244.

[83] Jennrich, R. I. (1969). "Asymptotic Properties of Nonlinear Least Squares Estimators." *Annals of Mathematical Statistics*, 40, 633–643.

[84] Jorgenson, B. (1983). "Maximum Likelihood Estimation and Large-Sample Inference for Generalized Linear and Nonlinear Regression." *Biometrika*, 70, 19–28.

[85] Judge, G. G. and Bock, M. E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimation in Econometrics*. Amsterdam: North-Holland.

[86] Judge, G. G., Griffiths, W. E., Hill, R. C., Lutkepohl, H., and Lee, T.-C. (1985). *The Theory and Practice of Econometrics*. 2nd ed. New York: John Wiley & Sons.

[87] Judge, G. G., Miyazaki, S., and Yancey, T. (1985). "Minimax Estimators for the Location Vectors of Spherically Symmetric Densities." *Econometric Theory*, 1, 409–417.

[88] Khan, M. S. and Ross, K. Z. (1977). "The Functional Form of the Aggregate Import Demand Equation." *Journal of International Economics*, 7, 149–160.

[89] Khazzoom, J. D. (1989). "A Note on the Application of the Nonlinear Two-Stage Least-Squares Estimator to a Box-Cox-Transformed Model." *Journal of Econometrics*, 42, 377–379.

[90] Kleijnen, J. P. C., Karremans, P. C. A., Oortwijn, W. K., and Van Groenendaal, W. J. H. (1987). "Jackknifing Estimated Weighted Least Squares: JEWLS." *Communications in Statistics — Theory and Methods*, 16, 747–764.

[91] Knight, J. L. (1986). "The Distribution of the Stein-Rule Estimator in a Model with Non-Normal Distributions." *Econometric Theory*, 2, 202–219.

[92] Lawrence, A. J. (1987). "A Note on the Variance of the Box-Cox Regression Transformation Estimate." *Applied Statistics*, 36, 221–223.

[93] Li, T. F. and Bhoj, D. S. (1988). "A Modified James-Stein Estimator with Application to Multiple Regression Analysis." *Scandinavian Journal of Statistics*, 15, 33–37.

[94] Lu, K. L. and Berger, J. O. (1989). "Estimation of Normal Means: Frequentist Estimation of Loss." *Annals of Statistics*, 17, 890–906.

[95] Maasoumi, E. (1978). "A Modified Stein-Like Estimator for the Reduced Form Coefficients of Simultaneous Equations." *Econometrica*, 46, 695–703.

[96] MacKinnon, J. G. and White, H. (1985). "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics*, 29, 305–325.

[97] Maddala, G. S. (1971). "Generalized Least Squares with an Estimated Variance Covariance Matrix." *Econometrica*, 39, 23–33.

[98] Magee, L. (1988). "The Behaviour of Modified Model When Some Values of the Dependent Variable Are Close to Zero." *The Review of Economics and Statistics*, 70, 362–366.

[99] Malinvaud, E. (1980). *Statistical Methods of Econometrics.* 3rd ed. Amsterdam: North-Holland.

[100] Mantel, N. (1987). "Understanding Wald's Test for Exponential Families." *The American Statistician*, 41, 147–148.

[101] Megbolugbe, I. F. (1986). "Econometric Analysis of Housing Trait Prices in a Third World City." *Journal of Regional Science*, 26, 533–547.

[102] Mills, T. C. (1978). "The Functional Form of the U. K. Demand for Money." *Applied Statistics*, 27, 52–57.

[103] Mittelhammer, R. C. (1985). "Quadratic Risk Domination of Restricted Least Squares Estimators via Stein-Ruled Auxiliary Constraints." *Journal of Econometrics*, 29, 289–303.

[104] Mizon, G. and Hendry, D. F. (1980). "An Empirical Application and Monte Carlo Analysis of Tests of Dynamic Specification." *Review of Economic Studies*, 57, 21–45.

[105] Montmarquette, C. and Bais, A. (1987). "A Survey Measures of Risk Aversion." *Economics Letters*, 25, 27–30.

[106] Morimune, K. (1989). "t Tests in a Structural Equation" *Econometrica*, 57, 1341–1360.

[107] Navidi, W. (1989). "Edgeworth Expansions for Bootstrapping Regression Models." *Annals of Statistics*, 17, 1472–1478.

[108] Nelson, F. D. and Savin, N. E. (1988). "The Nonmonotonicity of the Power Function of the Wald Test in Nonlinear Models." (University of Iowa, Department of Economics, Working paper No. 88-7).

[109] Norden, R. H. (1972a). "A Survey of Maximum Likelihood Estimation." *International Statistical Review*, 40, 329–354.

[110] Norden, R. H. (1972b). "A Survey of Maximum Likelihood Estimation — Part 2." *International Statistical Review*, 41, 39–58.

[111] Oberhofer, W. and Kmenta, J. (1974). "A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models." *Econometrica*, 42, 579–590.

[112] Orme, C. (1989). "On the Uniqueness of the Maximum Likelihood Estimator in Truncated Regression Models." *Econometric Review*, 8, 217–222.

[113] Patefield, W. M. (1985). "Information from the Maximum Likelihood Function." *Biometrika*, 72, 664–668.

[114] Peters, S. C. and Freedman, D. A. (1984). "Some Note on the Bootstrap in Regression Problems." *Journal of Business and Economic Statistics*, 2, 401–409.

[115] Phillips, P. C. B. (1977a). "Approximations to Some Finite Sample Distributions Associated with a First-Order Stochastic Difference Equation." *Econometrica*, 45, 463–485.

[116] Phillips, P. C. B. (1977b). "A General Theorem in the Theory of Asymptotic Expansions as Approximations to the Finite Sample Distributions of Econometric Estimators." *Econometrica*, 45, 1517–1534.

[117] Phillips, P. C. B. (1977c). "An Approximation to Finite Sample Distribution of Zellner's Seemingly Unrelated Regression Estimator." *Journal of Econometrics*, 6, 147–164.

[118] Phillips, P. C. B. (1980). "Finite Sample Theory and the Distribution of Alternative Estimators of the Marginal Propensity to Consume." *Review of Economic Studies*, 57, 183–224.

[119] Phillips, P. C. B. (1982). "Best Uniform and Modified Padé Approximants to Probability Densities in Econometrics," in *Advances in Econometrics*. Edited by W. Hildenbrand. Cambridge: Cambridge University Press.

[120] Phillips, P. C. B. (1983). "ERA's: A New Approach to Small Sample Theory." *Econometrica*, 51, 1505–1525.

[121] Phillips, P. C. B. (1984). "The Exact Distribution of the Stein-Rule Estimator." *Journal of Econometrics*, 25, 123–131.

[122] Phillips, P. C. B. (1988). "On the Formulation of Wald Tests of Nonlinear Restrictions." *Econometrica*, 56, 1065–1083.

[123] Poirier, D. J. (1978). "The Use of the Box-Cox Transformation in Limited Dependent Variable Models." *Journal of the American Statistical Association*, 73, 284–287.

[124] Poirier, D. J. and Melino, A. (1978). "A Note on the Interpretation of Regression Coefficients within a Class of Truncated Distributions." *Econometrica*, 46, 1207–1209.

[125] Poirier, D. J. and Ruud, P. A. (1979). "A Simple Lagrange Multiplier Test for Lognormal Regression." *Economics Letters*, 4, 251–255.

[126] Rayner, R. K. (1989). "Bootstrap Inversion of Edgeworth Expansions for Nonparametric Confidence Intervals." *Statistics and Probability Letters*, 8, 201–206.

[127] Richardson, D. H. and Rohr, R. J. (1971). "The Distribution of a Structural t-Statistic for the Case of Two Included Endogenous Variables." *Journal of the American Statistical Association*, 66, 375–382.

[128] Robert, C. (1988). "An Example Formula for the Risk of the Positive-Part James-Stein Estimator." *The Canadian Journal of Statistics*, 16, 161–168.

[129] Rocke, D. M. (1989). "Bootstrap Bartlett Adjustment in Seemingly Unrelated Regression." *Journal of the American Statistical Association*, 84, 598–601.

[130] Rothenberg, T. J. (1982). "Approximating the Distributions of Econometric Estimators and Test Statistics," in *Handbook of Econometrics*. Edited by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland.

[131] Rothenberg, T. J. (1984). "Hypothesis Testing in Linear Models When the Error Covariance Matrix Is Nonscalar." *Econometrica*, 52, 827–842.

[132] Royden, H. L. (1988). *Real Analysis*. New York: Macmillan Publishing.

[133] Ruppert, D. and Aldershof, B. (1989). "Transformations to Symmetry and Homoskedasticity." *Journal of the American Statistical Association*, 84, 437–446.

[134] Sargan, J. D. (1975). "Gram-Charlier Approximations Applied to t Ratios of k-Class Estimators." *Econometrica*, 43, 327–346.

[135] Sargan, J. D. (1976). "Econometric Estimators and the Edgeworth Approximation." *Econometrica*, 44, 421–448.

[136] Sargan, J. D. and Satchell, S. E. (1986). "A Theory of Validity for Edgeworth Expansions." *Econometrica*, 54, 189–213.

[137] Schaefer, R. L. (1986). "Alternative Estimators in Logistic Regression When the Data Are Collinear." *Journal of Statistical Simulation*, 25, 75–91.

[138] Schaefer, R. L., Roi, L. D., and Wolfe, R. A. (1984). "A Ridge Logistic Estimator." *Communications in Statistics — Theory and Methods*, 13, 99–113.

[139] Schlesselman, J. (1971). "Power Families: A Note on the Box and Cox Transformation." *Journal of Royal Statistical Society*, ser B, 33, 307–311.

[140] Sclove, S. L., Morris, C., and Radhakrishnan, R. (1972). "Non-Optimality of Preliminary-Test Estimators for the Mean of a Multivariate Normal Distribution." *Annals of Statistics*, 43, 1481–1490.

[141] Seaks, T. G. and Layson, S. K. (1983). "Box-Cox Estimation with Standard Econometric Problems." *The Review of Economics and Statistics*, 65, 160–164.

[142] Shao, J. (1988). "Bootstrap Variance and Bias Estimation in Linear Models." *The Canadian Journal of Statistics*, 16, 371–382.

[143] Shiryayev, A. N. (1984). *Probability*. New York: Springer-Verlag.

[144] Schucany, W. R., Gray, H. L., and Owen, D. B. (1971). "On Bias Reduction in Estimation." *Journal of the American Statistical Association*, 66, 524–533.

[145] Smyth, D. J. and Dua, P. (1986). "Inflation, Unemployment and the Median Voter." *Economics Letters*, 22, 181–186.

[146] Spanos, A. (1986). *Statistical Foundations of Econometric Modelling.* Cambridge: Cambridge University Press.

[147] Spitzer, J. J. (1976). "The Demand for Money, the Liquidity Trap, and Functional Forms." *International Economic Review*, 17, 220–227.

[148] Spitzer, J. J. (1978). "A Monte Carlo Investigation of the Box-Cox Transformation in Small Samples." *Journal of the American Statistical Association*, 73, 488–495.

[149] Spitzer, J. J. (1982a). "A Primer on Box-Cox Estimation." *The Review of Economics and Statistics*, 64, 307–313.

[150] Spitzer, J. J. (1982b). "A Fast and Efficient Algorithm for the Estimation of Parameters in Models with the Box-and-Cox Transformation." *Journal of the American Statistical Association*, 77, 760–766.

[151] Spitzer, J. J. (1984). "Variance Estimates in Models with the Box-Cox Transformation: Implications for Estimation and Hypothesis Testing." *The Review of Economics and Statistics*, 66, 645–652.

[152] Taylor, J. M. G. (1985). "Power Transformation to Symmetry." *Biometrika*, 72, 145–152.

[153] Taylor, W. E. (1983). "On the Relevance of Finite Sample Distribution Theory." *Econometric Review*, 2, 1–39.

[154] Teekens, R. and Koerts, J. (1972). "Some Statistical Implications of the Log Transformation of Multiplicative Models." *Econometrica*, 40, 793–819.

[155] Tse, Y. K. (1984). "Edgeworth Approximations for t-Ratios of 2SLS Estimates of a Dynamic Model." *Communications in Statistics — Simulation & Computation*, 13, 603–618.

[156] Ullah, A. (1974). "On the Sampling Distribution of Improved Estimators for Coefficients in Linear Regression." *Journal of Econometrics*, 2, 143–150.

[157] Ullah, A. (1982). "The Approximate Distribution Function of the Stein-Rule Estimator." *Economics Letters*, 10, 305–308.

[158] Ullah, A., Srivastava, V. K., and Chandra, R. (1983). "Properties of Shrinkage Estimator in Linear Regression When Disturbances Are Not Normal." *Journal of Econometrics*, 21, 389–402.

[159] Vaeth, M. (1985). "On the Use of Wald's Test in Exponential Families." *International Statistical Review*, 53, 199–214.

[160] Vinod, H. D. and Ullah, A. (1981). *Recent Advances in Regression Methods.* New York: Marcel Dekker.

[161] Wallace, D. L. (1958). "Asymptotic Approximations to Distributions." *Annals of Mathematical Statistics*, 29, 635–654.

[162] White, H. (1982). "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, 50, 1–25.

[163] White, H. (1984). *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

[164] White, K. J. (1972). "Estimation of the Liquidity Trap with a Generalized Functional Form." *Econometrica*, 40, 193–199.

[165] Wilks, S. S. (1962). *Mathematical Statistics*. New York: John Wiley & Sons.

[166] Zarembka, P. (1969). "Functional Form in the Demand for Money." *Journal of the American Statistical Association*, 63, 502–511.

[167] Zarembka, P. (1974). "Transformation of Variables in Econometrics," in *Frontiers in Econometrics*. Edited by P. Zarembka. New York: Academic Press.

VITAE

Minbo Kim was born in Kwangju, Korea on December 2, 1955, the second son of Mr. Young Mun Kim and Mrs. Youngja Chey. He received his elementary school education in Kwangju, graduating from Bo Sung Senior High school in Seoul, Korea. He entered Seoul National University in Seoul, Korea in 1975, and received his Bachelor of Business Administration in 1980. From 1980 to 1981, he served as a research associate at Korea Research Institute for Human Settlements. Kim was employed as a computer programmer/systems analyst at Honam Oil Refinery Co. in Seoul, Korea, from 1981 and 1984. In August of 1984, he entered the graduate school of LSU in Baton Rouge, Louisiana, and received the degree of Master of Science in Economics in May of 1988. He continued his graduate studies at LSU, working toward the degree of Doctor of Philosophy in Economics. During the course of his studies, he majored in Economics and minored in Mathematics. Youngja Shim is his wife and their daughter is Sun Kim.

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Minbo Kim

Major Field: Economics

Title of Dissertation: Small Sample Properties of Estimators and Test Statistics in Nonlinear Regression: The Box-Cox Transformation

Approved:

R. Carter Hill

Major Professor and Chairman

Dean of the Graduate School

EXAMINING COMMITTEE:

John C. Hilke

Stephen W. Fuerring

W. Douglas McMillin

D. Randolph Peterson

David J. Smyth

E. Jane Luzar

Date of Examination: July 30, 1990