Appeared in: Test, 1999, 8, 167-190.

Estimating a transformation and its effect on Box-Cox T-ratio

Zhenlin Yang
Department of Statistics and Applied Probability
c/o Department of Economics
National University of Singapore
10 Kent Ridge Crescent
Singapore 119260

Abstract

This article concerns i) the stochastic behavior of the Box-Cox transformation estimator and ii) the effect of estimating a transformation on the Box-Cox T-ratio used for the post-transformation analysis. It is shown that the transformation estimator depends on three factors: the model structure, the mean-spread and the error standard deviation σ_0. In general, a structured model is able to estimate the transformation very well; an unstructured model can do well also unless the mean-spread and σ_0 are both small; and a one-mean mode can give a poor estimate if σ_0 is small. When the sample is not large, it is shown that the unconditional effect of estimating a transformation on the Box-Cox T-ratio is generally small, and the 'conditional' effect is also negligible in most of the situations except the case of one-way ANOVA with small σ_0. Extensive Monte Carlo simulations are performed to support the theoretical findings.

Short title: Box-Cox T-ratio

Key Words: Asymptotic expansion, Box-Cox transformation, λ-fixed, Sensitivity, T-ratio.

AMS subject classification: 62F25

Appeared in: Test, 1999, 8, 167-190.

1 INTRODUCTION

In many applications of statistical modeling, a transformation of the dependent variable is required to achieve a normal theory linear model with a simple mean structure and homoscedastic errors. When such a transformation is known, the usual normal-theory linear model inference methods can be directly applied to the transformed responses. When the transformation is unknown, the common practice, as suggested by Box and Cox (1964), is to estimate the unknown transformation parameter and then select a nearest simple number corresponding to a *log* or *square root*, etc., transformation, and then carry out usual inferences for the parameters defined and interpreted on the selected scale.

Let $y = (y_1, \cdots, y_n)'$ be the vector of responses, and $h(y, \lambda) = [h(y_1, \lambda), \cdots, h(y_n, \lambda)]'$ the vector of transformed responses, where $h(\cdot, \lambda)$ is a strictly increasing transformation function, known except the transformation parameter λ, taking values on real line. Assume that there exists a true value λ_0 of λ such that the vector $h(y, \lambda_0)$ of the transformed observations satisfies

$$h(y, \lambda_0) = X\beta_0 + \sigma_0 e, \tag{1.1}$$

Where β_0 is a $p \times 1$ vector of regression parameters,σ_0 is the standard deviation of the error term, X is a known $n \times p$ matrix of full rank, and $\sigma_0 e$ is an $n \times 1$ vector of independent errors of same distribution.

Denote the parameter vector $(\beta_0', \lambda_0, \sigma_0)$' by ξ_0 and its estimator $(\hat{\beta}_n', \hat{\lambda}_n, \hat{\sigma}_n)'$ by $\hat{\xi}_n$. The restricted estimator of (β_0, σ_0) when λ_0 is known is denoted by $(\hat{\beta}_{n0}, \hat{\sigma}_{n0})$. Thus, when λ_0 is known, the post-transformation inference concerns β_0 and is carried out based on the λ_0-known T-ratio

$$T_0 = \frac{\sqrt{n}(\hat{\beta}_{n0} - \beta_0)}{\hat{\sigma}_{n0}}$$

which, after a suitable normalization, has a multivariate T-distribution when errors are exactly normal.

When λ_0 is unknown and is estimated by $\hat{\lambda}_n$, Box-Cox's analysis can be viewed as $\hat{\lambda}_n$ -fixed inference for $\beta_u(\hat{\lambda}_n)$, defined and interpreted on the selected scale $\hat{\lambda}_n$, based on the Box-Cox T-ratio

$$T_{BC}(\hat{\lambda}_n) = \frac{\sqrt{n}[\hat{\beta}_n - \beta_u(\hat{\lambda}_n)]}{\hat{\sigma}_n}$$

with the $\hat{\lambda}_n$ -fixed distribution of $T_{BC}(\hat{\lambda}_n)$ approximated by the distribution of T_0. For example, for a particular data set if the resulted estimate of the transformation parameter is $\hat{\lambda}_n = 0.5$, then Box and Cox fit the model $h(y, 0.5) = X\beta_u(0.5) + \sigma(0.5)e(0.5)$, and make inference about $\beta_u(0.5)$ by approximating the distribution of $T_{BC}(0.5)$ by that of T_0. Hinkley and Runger (1984, Sec.

2.1) and Carroll and Ruppert (1988, Sec. 4.3.4) gave a similar interpretation. Notice that $\hat{\lambda}_n$ can be equivalently replaced by a rounded value provided that rounding is done with reference to the confidence interval.

Questions arise as how much the Box-Cox T-ratio $T_{BC}(\hat{\lambda}_n)$ differs from the λ_0-known T-ratio T_0, and to what extent the $\hat{\lambda}_n$-fixed distribution of $T_{BC}(\hat{\lambda}_n)$ can be approximated by the distribution of T_0. These are the crucial questions to the validity of Box-Cox transformation methodology and are termed in this paper as the effect of estimating a transformation on the Box-Cox T-ratio. The former corresponds to the unconditional effect and the latter the conditional effect with $\hat{\lambda}_n$ regarded as fixed. Yang (1996) has studied these questions for large n, which lead to the asymptotic validity of the Box-Cox transformation methodology. In this article, we investigate these questions for small n case via a second-order asymptotic expansion of $T_{BC}(\hat{\lambda}_n)$. As this second-order expansion has a leading term T_0 and a smaller order 'affecting' term that involves $\hat{\lambda}_n$, it is necessary to investigate first the stochastic behavior of $\hat{\lambda}_n$, which is done by Yang (1997) and reexamined in this article with improved and extended results.

The above two problems (in short, behavior of$\hat{\lambda}_n$ and effect of $\hat{\lambda}_n$ on $T_{BC}(\hat{\lambda}_n)$ that will be studied in this article are closely related (directly or indirectly) to the two problems raised in Box and Cox (1982):

> *A.* There are numerous aspects of transformations that merit further study. These include in particular the further development of simple ways of assessing transformation potential; that is, of providing some formal measure of the ability of particular data to provide useful information about a class of transformations.
>
> *B.* Suppose that the parameter of interest (difference, regression coefficient, etc.) is defined on the data-dependent scale $\hat{\lambda}_n$; in what circumstances do confidence intervals for these parameters calculated in the "usual" way, as if $\hat{\lambda}_n$ were preassigned, provide an adequate approximation?

Section 2 presents general asymptotic expansions based on an M-estimation framework, followed by a specialization to the Box and Cox (1964) maximum likelihood estimation framework, which will be used throughout the article. Section 3 concerns the stochastic behavior of $\hat{\lambda}_n$. Section 4 studies the unconditional behavior of $T_{BC}(\hat{\lambda}_n)$. Section 5 investigates the $\hat{\lambda}_n$-fixed behavior of $T_{BC}(\hat{\lambda}_n)$. Each of the Sections 3 to 5 is accompanied by Monte Carlo results to back up the theoretical conclusions.

Putting $\eta = X\beta_0$, we now summarize the major conclusions and discuss their relations and implications to problems A and B. Most of the conclusions about $\hat{\lambda}_n$ were already reported in Yang (1997).

First, the stochastic behavior of $\hat{\lambda}_n$ depends on three factors: the model structure, the

spread in means ($\eta_i's$) and the error standard deviation σ_0. In general, structured models such as regression models or ANOVA models with at least two factors, are able to estimate λ_0 very well; the unstructured models, such as single factor ANOVA model, are able to estimate λ_0 well unless the spread in $\eta_i's$ and σ_0 are both small; and a one-mean model can also do well unless σ_0 is small.

The practical implications of above conclusions as related to problem A are as follows. A data set that came from an experiment using structured model is generally of high potential in determining the transformation. A data set that came from an experiment using unstructured or one-mean model still possesses a good potential in determining the transformation if the data stretch to a wide range relatively, otherwise it will be difficult to estimate the transformation.

As for the effect of estimation transformation on Box-Cox T-ratio, we find that, when n is small, the difference between $T_{BC}(\hat{\lambda}_n)$ and T_0 is small in general, and hence the distribution of $T_{BC}(\hat{\lambda}_n)$ can be well approximated by that of T_0. We also find that the $\hat{\lambda}_n$-fixed distribution of $T_{BC}(\hat{\lambda}_n)$ can be well approximated by the distribution of T_0 for all models when σ_0 is not small. When σ_0 is small, the approximation is still good in one-mean models and also reasonable in structured models if the fixed-$\hat{\lambda}_n$ is within two standard deviations of λ_0; in unstructured models, the $\hat{\lambda}_n$-fixed variance of ith element of $T_{BC}(\hat{\lambda}_n)$ can be deflated or inflated depending on the signs of $\eta_i - \bar{\eta}$ and $\hat{\lambda}_n - \lambda_0$, with the magnitude depending on $\eta_i - \bar{\eta}$, but the sum of $\hat{\lambda}_n$-fixed variances of the elements of $T_{BC}(\hat{\lambda}_n)$ is stable.

The implication of these conclusions for problem B is quite clear: whenever the $\hat{\lambda}_n$-fixed distribution of $T_{BC}(\hat{\lambda}_n)$ can be well approximated by the distribution of T_0 , then the usual confidence intervals for the $\hat{\lambda}_n$-dependent parameters will perform well. In this sense, all the $\hat{\lambda}_n$-fixed confidence intervals will perform well or reasonably well except the t-interval for the individual mean of a one-way ANOVA model with σ_0 small relative to the mean-spread.

Hooper and Yang (1997) studied problem B where they interpreted the Box-Cox method of post-transformation inferences as conditional inferences for $\beta_u(\hat{\lambda}_n)$ based on $T_{BC}(\hat{\lambda}_n)$ with the conditional distribution of $T_{BC}(\hat{\lambda}_n)$ given $\hat{\lambda}_n$ approximated by that of T_0. Yang (1996) showed under mild conditions that $T_{BC}(\hat{\lambda}_n)$ is asymptotically equivalent to T_0 and independent of $\hat{\lambda}_n$. Hence the two interpretations about the Box-Cox transformation methodology are asymptotically equivalent.

Bickel and Doksum (1981) argued that the inference should be unconditional about β_0 . They showed that the usual normal-theory inference methods can fail because of the variance inflation due to transformation estimation. Box and Cox (1982) commented that this variance inflation is obvious but irrelevant for any sensible scientific question. Hinkley and Runger (1984)

and Cox and Reid (1987) further supported the Box and Cox's approach by claiming that the slope parameters are stable in the so called z-scale that stabilizes or orthogonalizes the parameters. Duan (1993) showed that this claim is true under certain symmetric conditions on the regressors, but might fail when the symmetric conditions are not satisfied. Since the Box-Cox T-ratio is typically invariant under the z-transformation, the z-scale is not considered here. A review of the Box-Cox transformation technique is given by Sakia (1992).

2 ASYMPTOTIC EXPANSIONS

To facilitate the expansions, we introduce formally the definition of the parameter of interest after transformation selection. We employ the definition given by Cohen and Sackrowitz (1987), namely,

$$\beta_u(\hat{\lambda}_n) = \text{Expectation of } \hat{\beta}_n \text{ treating } \hat{\lambda}_n \text{as fixed.} \tag{2.1}$$

This definition is consistent with our $\hat{\lambda}_n$-fixed interpretation of the Box-Cox's post-transformation analysis. There are other definitions, e.g., $\beta_c(\hat{\lambda}_n)=E\{\hat{\beta}_n|\hat{\lambda}_n\}$ (Hinkley and Runger, 1984), which is asymptotically equivalent to (2.1) (Bickel, 1984). See also Yang (1992, Chapter 2) for a general discussion of the definition and interpretation of the parameter of interest following a transformation selection.

Consider first the general M-estimation framework, i.e.,$\hat{\xi}_n$ is an M-estimator of ξ_0 which solves

$$n^{-1}\sum_{i=1}^{n}\mathbf{\Psi}(y_i;\hat{\xi}_n) = 0_{(p+2)\times 1} \tag{2.2}$$

where $\mathbf{\Psi}_i$ is a $p+2$ dimensional vector-valued function having three components $\mathbf{\Psi}_{1i}$, $\mathbf{\Psi}_{2i}$ and $\mathbf{\Psi}_{3i}$ that correspond to β, λ_0 and σ_0 respectively. Thus $\mathbf{\Psi}_{2i}$ is dropped when λ_0 is known. Let

$$\bar{\mathbf{\Psi}} = \bar{\mathbf{\Psi}}(y,\xi_0) = n^{-1}\textstyle\sum_{i=1}^{n}\mathbf{\Psi}_i(y_i,\xi_0), \quad \dot{\mathbf{\Psi}} = \dot{\mathbf{\Psi}}(y,\xi_0) = (\partial/\partial\xi_0')\bar{\mathbf{\Psi}}(y,\xi_0),$$
$$\ddot{\mathbf{\Psi}} = \ddot{\mathbf{\Psi}}(y,\xi_0) = (\partial/\partial\xi_0)\dot{\mathbf{\Psi}}(y,\xi_0), \qquad \mathbf{A} = \mathrm{E}\dot{\mathbf{\Psi}}, \qquad \mathbf{B} = \mathrm{E}\ddot{\mathbf{\Psi}},$$

where $\bar{\mathbf{\Psi}}$ is a $(p+2)\times 1$ vector, $\dot{\mathbf{\Psi}}$ and $\mathbf{A}$ are $(p+2)\times(p+2)$ matrices, $\ddot{\mathbf{\Psi}}$ and $\mathbf{B}$ are $(p+2)^2\times(p+2)$ matrices or $(p+2)\times(p+2)\times(p+2)$ arrays. They are all partitioned according to $(\beta_0,\lambda_0,\sigma_0)$. The subvectors of $\bar{\mathbf{\Psi}}$ are denoted by $\bar{\mathbf{\Psi}}_1$, $\bar{\mathbf{\Psi}}_2$ and $\bar{\mathbf{\Psi}}_3$, the submatrices of $\dot{\mathbf{\Psi}}$ and $\mathbf{A}$ by $\dot{\mathbf{\Psi}}_{ij}$ and $\mathbf{A}_{ij}$, $i,j=1,2,3$, and the subarrays of $\ddot{\mathbf{\Psi}}$ and $\mathbf{B}$ by $\ddot{\mathbf{\Psi}}_{ijk}$ and $\mathbf{B}_{ijk}$, $i,j,k=1,2,3$. For example,

$$\begin{aligned}
\bar{\mathbf{\Psi}}_1 &= \bar{\mathbf{\Psi}}_1(y,\xi_0 &&= n^{-1}\textstyle\sum_{i=1}^{n}\mathbf{\Psi}_{1i}(y_i,psi_0), && \text{a } p\times 1\text{vector},\\
\dot{\mathbf{\Psi}}_{11} &= \dot{\mathbf{\Psi}}_{11}(y,\xi_0 &&= (\partial/\partial\beta_0')\bar{\mathbf{\Psi}}_1(y,\psi_0), && \text{a } p\times p\text{matrix},\\
\ddot{\mathbf{\Psi}}_{111} &= \ddot{\mathbf{\Psi}}_{111}(y,\xi_0) &&= (\partial^2/\partial\beta_0\partial\beta_0')\bar{\mathbf{\Psi}}_1(y,\xi_0), && \text{a } p^2\times p\text{matrix or a} p\times p\times p\text{array}.
\end{aligned}$$

Note that all the quantities introduced above depend on n implicitly. Assume

$C1$. $\hat{\xi}_n$ is root-n consistent,

$C2$. $\bar{\mathbf{\Psi}} = O_p(n^{-1/2})$, and $\mathrm{E}\bar{\mathbf{\Psi}} = 0$,

$C3$. $\dot{\mathbf{\Psi}} = \mathbf{A} + O_p(n^{-1/2})$,

$C4$. $\ddot{\mathbf{\Psi}} = \mathbf{B} + O_p(n^{-1/2})$,

$C5$. $\mathbf{A}$ and $\mathbf{A^{-1}}$ are $O(1)$, with $\mathbf{A_{13}} = \mathbf{A}'_{31} = O(n^{-1/2})$.

Assume further that the remainder term in the second-order Taylor expansions of the elements of $\mathbf{\Psi_i}$ has the order of $(\hat{\xi_n} - \xi_0)^3$, and that a random quantity bounded in probability has a finite expectation. Now we present some general results. The proofs are tedious and sketches are given in the Appendix.

Theorem 2.1. *Under the assumptions C1-C5,if* $\beta_u(\lambda_0) = \beta_0 + O(n^{-3/2})$, *then as* $n \to \infty$, *we have a second-order asymptotic expansion for* $T_{BC}(\hat{\lambda}_n)$ *and a first-order expansion for* $\hat{\lambda}_n$,

$$T_{BC}(\hat{\lambda}_n) = T_0 + U_1(\hat{\lambda}_n - \lambda_0) + U_2(\hat{\lambda}_n - \lambda_0)^2 + O_p(n^{-1}, \tag{2.3}$$

$$\hat{\lambda}_n - \lambda_0 = \frac{\bar{\mathbf{\Psi}}_2\mathbf{A}_{23}\mathbf{A}_{33}^1\bar{\mathbf{\Psi}}_3 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\bar{\mathbf{\Psi}}_1}{\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} - \mathbf{A}_{22} + \mathbf{A}_{23}\mathbf{A}_{33}^{-1}\mathbf{A}_{32}} + O_p(n^{-1}), \tag{2.4}$$

where U_1 *and* U_2 *in (2.3) are both* $O_p(1)$ *with the detailed expressions given at the end of Appendix.*

It is unorthodox to keep a term $U_2(\hat{\lambda}_n - \lambda_0)^2$ that is of the same order as the error term $O_p(n^{-1})$ in the expression. However, for a fixed n, certain approximations (see Sections 4 and 5) show that, for a structured model, as $\sigma_0 \to 0$, $U_2(\hat{\lambda}_n - \lambda_2(\hat{\lambda}_n - \lambda_0)^2 = O_p(1)$ whereas $U_1(\hat{\lambda}_n - \lambda_0) = O_p(\sigma_0)$, showing that the magnitude of $U_2(\hat{\lambda}_n - \lambda_0)^2$ will exceed that of $U_1(\hat{\lambda}_n - \lambda_0$ as σ_0 goes to small. Hence this term is important for studying the small-σ_0 behavior of $T_{BC}(\hat{\lambda}_n)$. This term vanishes for unstructured models.

The expansion (2.3) indicates that $T_{BC}(\hat{\lambda}_n)$ and T_0 differ only on second order, hence they are asymptotically equivalent and the Box-Cox $\hat{\lambda}_n$-fixed inference is asymptotic valid. Yang (1996) reached the same conclusion using the first-order expansion of $T_{BC}(\hat{\lambda}_n)$. He also gave a first-order expansion for the Bickel-Doksum T-ratio, obtained by replacing $\beta_u(\hat{\lambda}_n)$ in $T_{BC}(\hat{\lambda}_n)$ by β_0, which indicates that the Bickel-Doksum T-ratio differs from T_0 even on the first order, hence the unconditional inference for β_0 by approximating the distribution of Bickel-Doksum T-ratio by that of T_0 is not valid. This agrees with the observations made by Bickel and Doksum (1981).

Although the Box-Cox's $\hat{\lambda}_n$-fixed inference is asymptotically valid, its performance for moderate sample sizes is still unclear to us. Theorem 2.1 provides a tool for tackling this problem. Notice that in developing Theorems 2.1, we have assumed that $\mathbf{A}_{13} = O(n^{-1/2})$. This is not restrictive since $\mathbf{A}_{13}$ corresponds to β_0 and σ_0 which are, respectively, the location and scale parameters of the transformed model hence are orthogonal in the sense of Cox and Reid (1987).

In the cases that errors are exactly normal and maximum likelihood estimation method is used we have $\mathbf{A}_{13} = 0$. There is no difficulty theoretically in deriving the asymptotic expansions if this assumption is dropped, but the derivation will be more tedious.

Theorem 2.1 corresponds to the general M-estimation framework hence $\mathbf{\Psi}_i$ and h functions need not be specified, as long as the assumptions C1-C5 are satisfied. To study the stochastic behavior of $\hat{\lambda}_n$ and $T_{BC}(\hat{\lambda}_n)$ in detail, it is necessary to specify the functions $\mathbf{\Psi}_i$ and h. Clearly, the Box-Cox power transformation and the score function under normal errors are the popular candidates. In this case, we have $h(t,\lambda) = (t^\lambda - 1)/\lambda$, if $\lambda \neq 0$; and $\log t$, if $\lambda = 0$, and

$$\mathbf{\Psi}_i(y_i,\xi_0) = \begin{cases} \mathbf{\Psi}_{1i}(y_i,\xi_0) & = \sigma_0^{-2} x_i[h(y_i,\lambda_0) - x_i'\beta_0], \\ \mathbf{\Psi}_{2i}(y_i,\xi_0) & = \log y_i - \sigma_0^{-2}[h(y_i,\lambda_0) - x_i'\beta_0]\dot{h}(y_i,\lambda_0), \\ \mathbf{\Psi}_{3i}(y_i,\xi_0) & = \sigma_0^{-3}[h(y_i,\lambda_0) - x_i'\beta_0]^2 - \sigma_0^{-1}, \end{cases} \tag{2.5}$$

where $\dot{h}(y_i,\lambda_0) = (\partial/\partial\lambda_0)h(y_i,\lambda_0)$.

The estimators from (2.5) are usually called the **Box-Cox estimators**. In this article, we will concentrate on these. The estimators corresponding to other $\mathbf{\Psi}_i$ and h functions can be studied in a similar way. Now, let I_n denote an $n \times n$ identity matrix and $Q = I_n - X(X'X)^{-1}X'$, and let $\dot{h} = \dot{h}(y_i,\lambda_0)_{n\times i}$ and $\ddot{h} = (\partial^2/\partial\lambda_0^2)h(y_i,\lambda_0)_{n\times 1}$. Theorem 2.1 can be easily reduced to the following.

Corollary 2.1. *Assume that the $\mathbf{\Psi}_i$ function in (2.5) satisfies the assumptions C1-C5. Then,*

$$T_{BC}(\hat{\lambda}_n) = T_0 + U_1(\hat{\lambda}_n - \lambda_0) + U_2(\hat{\lambda}_n - \lambda_0)^2 + O_p(n^{-1}), \tag{2.6}$$

$$\frac{\hat{\lambda}_n - \lambda_0}{\sigma_0} = \frac{\sigma_0 \sum_{i=1}^n \log y_i + e'(P\mathrm{E}\dot{(h)} - \dot{h}) - (1 - e'e/n)\mathrm{E}(e'\dot{(h)})}{\mathrm{E}(\dot{h}'\dot{h}) - \mathrm{E}(\dot{h}')P\mathrm{E}(\dot{h}) + \sigma_0\mathrm{E}(e'\ddot{h}) - 2n^{-1}[\mathrm{E}(e'\dot{h})]^2} \tag{2.7}$$

where $U_1 = \sqrt{n}\sigma_0^{-1}(X'X')^{-1}X'[\dot{h} - \mathrm{E}\dot{h} - eE(e'\dot{h})/n]$ and $U_2 = -(2\sigma_0^2\sqrt{n})^{-1}(X'X)^{-1}X'eE(\dot{h}'Q\dot{h})$.

Proof. For (2.6), evaluate all the quantities involved in U_1 and U_2 of (2.3) using (2.5) and the expressions of the Box-Cox estimators $\hat{\sigma}_n^2 = n^{-1}[h'(y,\hat{\lambda}_n)Qh(y,\hat{\lambda}_n)]$ and $\hat{\sigma}_{n0}^2 = n^{-1}[h'(y,\lambda_0)Q$ $h(y,\lambda_0)]$. Then eliminate the terms that are either $O_p(n^{-1}$ or negligible when σ_0 is small. For (2.7), evaluate all the quantities involved in (2.4) and simplify. Clearly, the third term in (2.6) is $O_p(n^{-1})$, negligible with respect to n but may not be negligible with respect to σ_0 when it is small since σ_0^{-2} is involved in U_2.

With the simplified results of Corollary 2.1, it is possible to study in detail the stochastic behavior of $\hat{\lambda}_n$ (Section 3) and its effect on the Box-Cox T-ratio (Sections 4 and 5). When $\lambda_0 = 0$, it is possible to express (2.6) and (2.7) explicitly in terms of X,σ_0 and e. However, when $\lambda_0 \neq 0$ it is necessary to introduce a further approximation to *logy*.

Appeared in: Test, 1999, 8, 167-190.

3 ESTIMATION OF BOX-COX TRANSFORMATION

This section presents detailed results regarding the stochastic behavior of the Box-Cox estimator $\hat{\lambda}_n$ in various situations by further evaluating or approximating (2.7). Yang (1997) also obtained the expansion (2.7) that was then approximated by the small σ_0 method. His results are improved and extended by considering separately the log-transformation ($\lambda_0 = 0$) where no further approximation is necessary and other transformations ($\lambda_0 \neq 0$) where a better approximation method is used. We first present some theoretical results and then some Monte Carlo simulations. Throughout this article, we use # to denote the elementwise vector multiplication. Common functions such as log applied to a vector are operated elementwise. A vector subtracted by a scalar means elementwise subtraction.

3.1 Theoretical Results

Theorem 3.1. *Assuming the first six moments of e_1 are the same as those of a standard normal random variable, when $\lambda_0 = 0$, we have for large n,*

$$\frac{\hat{\lambda}_n - \lambda_0}{\sigma_0} = \frac{-\frac{1}{2}(Q\eta^2)'e - \sigma_0(\eta - \bar{\eta})'e^2 + \frac{1}{2}\sigma_0^2 1_n'(3e - e^3)}{\frac{1}{4}\|Q\eta^2\|^2 + 2\sigma_0^2\|\eta - \bar{\eta}\|^2 + \frac{3}{2}n\sigma_0 4} + O_p(n^{-1}), \tag{3.1}$$

$$Var(\hat{\lambda}_n) = \frac{\sigma_0^2}{\frac{1}{4}\|Q\eta^2\|^2 + 2\sigma_0^2\|\eta - \bar{\eta}\|^2 + \frac{3}{2}n\sigma_0^4} + O_p(n^{-1}); \tag{3.2}$$

when $\lambda_0 \neq 0$,letting $\theta_i = \lambda_0\sigma_0(1 + \lambda_0\eta_i)^{-1}, i = 1, \cdots, n$, we have for small θ_0 and large n,

$$\frac{\hat{\lambda}_n - \lambda_0}{\lambda_0} = \frac{-(\theta^{-1}\#\phi + \frac{1}{2}\theta)'Qe - (\phi - \bar{\phi})'e^2 + \frac{1}{2}\theta'(3e - e^3)}{\|Q(\theta^{-1}\#\phi + \frac{1}{2}\theta\|^2 + 2\|\phi - \bar{\phi}\|^2 + \frac{3}{2}\|\theta\|^2} + O_p(n^1) + O_p(\theta_0^3), \tag{3.3}$$

$$Var(\hat{\lambda}_n) = \frac{\lambda_0^2}{\|Q(\theta^1\#\phi + \frac{1}{2}\theta\|^2 + 2\|\phi - \bar{\phi}\|^2 + \frac{3}{2}\|\theta\|^2} + O_p(n^1) + O_p(\theta_0^3). \tag{3.4}$$

where $\theta_0 = \max|\theta_i|$, $\theta = \{\theta_i\}_{n\times 1}$, $\phi_i = \log(1 + \lambda_0\eta_i)$, $\phi = \{\phi_i\}_{n\times 1}$, $\bar{\phi} = n^{-1}\sum_{i=1}^n \phi_i$ and 1_n is a vector of 1's.

Proof. When $\lambda_0 = 0$, it is easy to show that $\dot{h}(y_i, 0) = \lim_{\lambda_0\to 0}\dot{h}(y_i, \lambda_0) = \frac{1}{2}(\log y_i)^2$ and $\ddot{h}(y_i, 0) = \lim_{\lambda_0\to 0}\ddot{h}(y_i, \lambda_0) = \frac{1}{2}(\log y_i)^3$. Substituting $X\beta_0 + \sigma_0 e$ into (2.7) for $\log y$ gives (3.1) and (3.2).

When $\lambda_0 \neq 0$, we have $\dot{h}(y_i, \lambda_0) = \lambda_0^{-1}[1 + \lambda_0 h(y_i, \lambda_0)]\log y_i - \lambda_0^{-1}h(y_i, \lambda_0)$ and $\ddot{h}(y_i, \lambda_0) = [\dot{h}(y_i, \lambda_0) - \lambda_0^{-2}]\log y_i + \lambda_0^{-2}h(y_i, \lambda_0) - \lambda_0^{-1}\dot{h}(y_i, \lambda_0)$. Thus, it is only necessary to further approximate $\log y_i$ to make (2.7) explicit. From the relation $\log y_i = \lambda_0^{-1}\log[1 + \lambda_0 h(y_i, \lambda_0)]$, we have by a Taylor expansion

$$\lambda_0 \log y_i = \log(1 + \lambda_0\eta_i) + \theta_i e_i - \frac{1}{2}\theta_i^2 e_i^2 + O_p(\theta_i^3). \tag{3.5}$$

Now using (3.5), some tedious algebraic work leads to (3.3) and (3.4). $\diamond$

The proof of the first part of the theorem can also be reached from the second part by letting $\lambda_0 \to 0$. The approximation (3.5) should be sufficient for most of the practical purposes as it is necessary that the θ_i's are small to guarantee positive y_i's. The small θ_i's can be achieved when at least one of the following conditions is satisfied: i) σ_0 is small, ii) $\eta_{(1)} = \min|\eta_i|$ is large, and iii) λ_0 is small. When only a first-order approximation is used, it is called the small-θ method by Draper and Cox (1969). The approximation (3.5) is more accurate than the small-θ method and more general than the small-σ approximations widely applied (Bickel and Doksum, 1981; Hooper and Yang, 1997; and Yang, 1997). Note that having i) or ii) is equivalent to having a large signal-to-noise ratio or a small coefficient of variation.

Theorem 3.1 improves and extends the results of Yang (1997). For $\lambda_0 = 0$, Bickel and Doksum (1981) reported explicit formulas for the cases of one-mean model, one-way layout and two-way layout with additive effects, which turn out to be the special cases of our formula (3.2). Other works regarding $\hat{\lambda}_n$ and $Var(\hat{\lambda}_n)$ include Draper and Cox (1969), Hinkley (1975), Atkinson (1985), Lawrance (1987) and Cox and Reid (1987, Sec. 5). Yang (1998) derived (3.4) in a non-rigorous manner.

Theorem 3.1 explicitly reveals the three factors governing the behavior of $\hat{\lambda_n}$, namely the model structure, the mean-spread and the error standard deviation, being respectively the first, second and third term in the denominator of (3.4). Following conclusions can easily be drawn: in general, structured models such as regression models or ANOVA models with at least two factors are able to estimate λ_0 very well; the unstructured models such as single factor ANOVA models are able to estimate λ_0 well unless the spread in η_i's and σ_0 are both small; and a one-mean model can also do well unless σ_0 is small.

When σ_0 is large, all the models perform in a similar way and estimation of transformation can be very easy. This point becomes clearer by observing that $Var(\hat{\lambda}_n)$ in (3.2) goes to 0 as $\sigma_0 \to \infty$, n fixed.

Finally, Theorem 3.1 also allows us to see the distributional property of $\hat{\lambda}_n$: $\hat{\lambda}_n$ is governed by three uncorrelated terms: the first is normal, and the second and third have zero means and are asymptotically normal under standard conditions. Hence when n is large, $\hat{\lambda}_n$ should possess a distribution quite close to normal. This point can not be clearly seen from the results of Yang (1997).

In summary, i) our results can be used to measure the transformation potential of a particular set of data (Box and Cox, 1982), i.e., the extent to which it is feasible to determine a suitable transformation from a particular type of data; ii) the results can be used for statistical inferences about λ_0, such as developing a new test of λ_0; and iii) the results make it much easier the study of

a more important problem of this article: the effect of estimating transformation on the Box-Cox T-ratio.

Cox and Reid (1989) derived an expression for $Var(\hat{\lambda}_n)$ based on an orthogonal parameter setting, which has a very similar structure as ours. It also involves three factors, namely the squared coefficient of variation from the regression component, the coefficient of variation of the error component, and a kind of signal to noise ratio. However, our expression is explicit in terms of the parameters λ_0, β_0, and σ_0, and the design X, so that the effect of each of the factors on $Var(\hat{\lambda}_n)$ can be seen clearly.

3.2 Monte Carlo Simulations

All the conclusions drawn from Theorem 3.1 rely on the assumption of large sample size n. It is necessary to check the applicability of these conclusions (i.e., the accuracy of the formulas) when n is not large. For various parameter configurations, the standard deviation (sd) of $\hat{\lambda}_n$ is simulated and compared with that calculated from (3.2) or (3.4). We consider the parameter configurations such that the probability $P[(1+\lambda_0(\eta_i+\sigma_0 e_i)) < 0]$ is negligible for all $i = 1, \cdots, n$. When $\lambda_0 = 0$, there is no restrictions theoretically, but numerically it is restrictive to have a very large mean. Each simulated number is based on 10,000 random samples. We consider three completely different models for illustration:

Model 1. A one-mean model with parameter settings: $n = 50$; $\lambda_0 = 0.0$, 0.05, 0.2 and 0.5; $\beta_0 = 0.1$, 1.0, 5.0, 10.0, and 20.0; $\sigma_0 = 0.1$, 1.0, 2.0, and 10.0.

Model 2. A three-means model. The parameter settings are: $n = 36$ (12 for each mean); $\lambda_0 = 0.0$ and 0.1; $\beta_0' = (31, 21, 11)$, (21, 11, 1), and (12, 11, 10); $\sigma_0 = 0.01$, 0.1, 1.0, 2.0, and 5.0.

Model 3. A 3^3 factorial design with linear effects where $n = 27$, $\beta_0' = (5.2523, 0.569, -0.4312, -0.2682)$, the fitted value of the textile example of Box and Cox (1964); $\lambda_0 = 0.0$, 0.1 and -0.05; and $\sigma_0 = 0.01$, 0.1, 1.0, 2.0, and 5.0.

The selected results are summarized in Table 3.1. Detailed results are available from the author upon request. The results generally show that the formulas (3.2) and (3.4) are very accurate. Thus, the analytical conclusions are applicable when n is not large. It is interesting to note that in the one-mean model the sd of $\hat{\lambda}_n$ decreases almost linearly with the increase of σ_0. In the three-means model, the sd of $\hat{\lambda}_n$ decreases significantly as the means move further apart and when the mean-spread is large the sd of $\hat{\lambda}_n$ depends very little on σ_0; when σ_0 is large relative to the mean-spread, the sd of $\hat{\lambda}_n$ is small. Furthermore, the magnitude of β_0 plays no role in the behavior of when $\lambda_0 = 0$, but plays some role when $\lambda_0 \neq 0$. In the 3^3 factorial design, the sd of $\hat{\lambda}_n$ reaches to its maximum at a certain value of σ_0.

Table 3.1: Selected results for the simulated and calculated (lower entry) $sd's$ of $\hat{\lambda}_n$

Model 1, $\lambda_0 = 0.0$						Model 2, $\lambda_0 = 0.1$				
$\sigma_0 \ddots \beta_0$	0.1	1.0	5.0	20.0		$\beta_0 \ddots \sigma_0$	0.01	0.1	1.0	2.0
0.1	1.2341	1.2402	1.2143	–		31,21,11	0.0471	0.0468	0.0446	0.0459
	1.1547	1.1547	1.1547	1.1547			0.0430	0.0430	0.0427	0.0419
1.0	0.1232	0.1227	0.1243	0.1237		21,11,1	0.0299	0.0298	0.0197	0.0294
	0.0115	0.0115	0.0115	0.0115			0.0276	0.0276	0.0274	0.0267
10.0	0.0124	0.0124	0.0122	0.0123		12,11,10	0.4586	0.4839	0.3392	0.2096
	0.0115	0.0115	0.0115	0.0115			0.4472	0.4448	0.3067	0.1906
Model 3, $\lambda_0 = 0.0$						Model 3, $\lambda_0 = -0.15$				
$\sigma_0 = 0.1$	0.1	1.0	2.0	5.0		$\sigma_0 = .01$	0.1	1.0	2.0	5.0
0.0083	0.0817	0.1472	0.0856	0.0359		0.0064	0.0601	0.1083	0.0618	–
0.0085	0.0792	0.1262	0.0739	0.0311		0.0063	0.0585	0.0928	0.0543	0.0229

4 THE UNCONDITIONAL EFFECT

We now start to investigate the effect of estimating a transformation on the Box-Cox T-ratio. We first study the simpler unconditional effect, and then the harder conditional effect with regarded as fixed in next section. Again the theoretical results are followed by the Monte Carlo simulations. Simulation serves the purpose of confirmation of using second-order expansion to approximate the small sample effect of transformation estimation, and thus checks the reliability of the theoretical conclusions.

4.1 Theoretical Results

Theorem 4.1. *Assuming i) the first six moments of e_1 are the same as those of a standard normal random variable, ii) $X'X = O(n)$, and iii) the conditions of Corollary 2.1 are true, we have*

$$T_{BC}(\hat{\lambda}_n) = T_0 + \tilde{U}_1(\hat{\lambda}_n - \lambda_0) + \tilde{U}_2(\hat{\lambda}_n - \lambda_0)^2 + O_p(n^{-1}) + O_p(\theta_0^3), \quad (4.1)$$

$$Var[T_{BC}(\hat{\lambda}_n)] = Var(T_0) + O_p(n^{-1}) + O_p(\theta_0^3), \quad (4.2)$$

where $\tilde{U}_1 = \sqrt{n}\lambda_0^{-1}(X'X)^{-1}X'[(\phi-\bar{\phi})\#e+\frac{1}{2}\theta\#(e^2-1)]$, *and* $\tilde{U}_2 = -\frac{1}{2}n^{1/2}(X'X)^{-1}X'e\|Q(\theta^{-1}\#\phi)\|^2$.

Proof. For (4.1), evaluating U_1 of (2.6) using (3.5) gives $\tilde{U}_1$. The transition from U_2 to $\tilde{U}_2$ is based on the following arguments. For fixed $\sigma_0, U_2(\hat{\lambda}_n-\lambda_0)^2 = O_p(n^{-1})$, which should be absorbed into the error term. However, for structured models with n fixed, it can be seen using (3.5) and the results of Theorem 3.1 that, as $\sigma_0 \to 0, U_2(\hat{\lambda}_n - \lambda_0)^2 = O_p(1)$ while $U_1(\hat{\lambda}_n - \lambda_0) = O_p(\sigma_0)$,

suggesting that the leading term of U_2, that is $\tilde{U}_2$, should be kept for the purpose of small σ_0 study. This term becomes more important when studying the $\hat{\lambda}_n$-fixed behavior of $T_{BC}(\hat{\lambda}_n)$ in Section 5.

For (4.2), based on the general assumption of Section 2 that a quantity bounded in probability has a finite expectation it suffices to show that $Cov[T_0, \tilde{U}_2(\hat{\lambda}_n - \lambda)^2]$ is of order $O_p(n^{-1})$ that is easy by Theorem 3.1 and that $Cov[T_0, \tilde{U}_2(\hat{\lambda}_n - \lambda_0)^2]$ is negligible when σ_0 is small. Some algebra leads to $Cov[T_0, \tilde{U}_2(\hat{\lambda}_n - \lambda_0)^2] \approx -2(X'X)^{-1}$ when σ_0 is small, showing that it is negligible. $\diamond$

Note that for the case of log transformation, the $O_p(\theta_0^3)$ term in Theorem 4.1 vanishes and the exact expressions of U_1 and U_2 can be obtained by either the relation $\dot{h}(y,0) = \frac{1}{2}\log^2 y$ directly or letting $\lambda_0 \to 0$ in the expressions of $\tilde{U}_1$ and $\tilde{U}_2$, which gives

$$U_1 = \sqrt{n}(X'X)^{-1}X'[(\eta - \bar{\eta})\#e + \frac{1}{2}\sigma_0(e^2 - 1)] \text{ and } U_2 = -(8\sigma_0^2\sqrt{n})^{-1}(X'X)^{-1}X'e\|Q\eta^2\|^2.$$

The results in Theorem 4.1, especially (4.2), is certainly encouraging. It says that when n is small, estimating the transformation has very little effect on the variance of $T_{BC}(\hat{\lambda}_n)$ in any situations. In contrast, $\hat{\lambda}_n$ depends very much on the model type, mean spread and magnitude of error variance. Our results also show that $T_{BC}(\hat{\lambda}_n)$ does not depend on the symmetry of design. In contrast, the stability (with respect to $\hat{\lambda}_n$) of the estimated slope coefficients on the z scale depends on the symmetry of design (Duan, 1993). The small n effect (unconditional) of estimating transformation on higher moments of $T_{BC}(\hat{\lambda}_n)$ can be studied in detail by examining the "affecting" term $U_1(\hat{\lambda}_n - \lambda_0) + U_2(\hat{\lambda}_n - \lambda_0)^2$ at various situations, which can be done in connection with Theorem 3.1. The investigations can be made easier by first concentrating on the log transformation (λ_0=0) and then generalizing to other transformations.

(i) When $Q\eta^2 \neq 0$, which happens when model for the means have structure, such as regression models or factorial models with two or more factors, the effect is small if σ_0 is small since $U_1\hat{\lambda}_n$ is small in the sense of σ_0 and $U_2\hat{\lambda}_n^2$ is small in the sense of n. This point becomes more evident from the limits:

$$\lim_{\sigma_0 \to 0}(U_1\hat{\lambda}_n) = 0 \text{ and } \lim_{\sigma_0 \to 0}(U_2\hat{\lambda}_n^2) = -\frac{1}{2}n^{-1/2}\|Q\eta^2\|^{-2}(X'X)^{-1}X'e[e'Q\eta^2]^2,$$

where the second one is clearly of the same order as the error term $O_p(n^{-1})$.

(ii) When $Q\eta^2 = 0$, which occurs when model for the means does not have structure, such as one factor ANOVA model, U_2 vanishes and $U_1\hat{\lambda}_n$ is a quantity of order $O_p(n^{-1/2})$. When σ_0 is small $U_1\hat{\lambda}_n$ is quite stable with respect to the changes in parameter values as seen from the limit,

$$\lim_{\sigma_0 \to 0}(U_1\hat{\lambda}_n) = \frac{1}{2}\sqrt{n}\|\eta - \bar{\eta}\|^{-2}(X'X)^{-1}X'[(\eta - \bar{\eta})\#e](\eta - \bar{\eta})'e^2$$

. (iii) When $\|\eta - \bar{\eta}\| = 0$, which happens when all means are equal, we have

$$U_1\hat{\lambda}_n \approx \frac{1}{6}n^{-3/2}1_n'(3e - e^3)1_n'(e^2 - 1),$$

a quantity independent of σ_0 and β_0, of order $O_p(n^{-1/2})$ and expected to be small. This is in contrast to the behavior of $\hat{\lambda}_n$ in one-mean models where $Var(\hat{\lambda}_n)$ can be very large when σ_0 is small.

(iv) Finally when σ_0 is large, all models behave like a one-mean model as evidenced by the limit

$$\lim_{\sigma_0\to\infty}(U_1\hat{\lambda}_n) = \frac{1}{6}n^{-3/2}1_n'(3e - e^3)1_n'(e^2 - 1),$$

hence the effect in this case is expected to be small in general.

The discussions i)-iii) above extend directly to $\lambda_0 \neq 0$ cases. The last discussion extends to $\lambda_0 \neq 0$ cases where σ_0 is large but θ_0 is still small.

4.2 Monte Carlo Simulations

We now present some simulation results to confirm our theoretical conclusions. We consider again the three models used in Section 3 with similar parameter configurations. The selected results are put in Tables 4.1. When the errors are exactly normal, $Var(T_0) = [n^2/(n - p - 3)](X'X)^{-1}$ with the degrees of freedom (df) of T_0 reduced by one, i.e., $df = n - p - 1$. The reduction in df is to account for the estimation of . Simulation results exhibit a general excellent agreement between $Var[T_{BC}(\hat{\lambda}_n)]$ and $Var(T_0)$. This illustrates the accuracy of second-order expansion (4.2) when n is not large.

To demonstrate the effect of design, we also consider an asymmetric 2^3 factorial design obtained by modifying the design matrix in the above symmetric 33 factorial design by changing the level '-1' to '0', while leaving the others unchanged (Duan, 1993). The results in Table 4.1 show that symmetry of design is not important to the behavior of $T_BC(\hat{\lambda}_n)$.

5 THE EFFECT OF TRANSFORMATION MISSPECIFICATION

We now study the effect of transformation misspecification on the Box-Cox T-ratio, i.e., the $\hat{\lambda}_n$-fixed behavior of $T_{BC}(\hat{\lambda}_n)$. That is the key issue to the validity of the Box-Cox transformation methodology. This is done by comparing the $\hat{\lambda}_n$-fixed mean and variance of $T_{BC}(\hat{\lambda}_n)$ with the mean and variance of T_0. The study of the $\hat{\lambda}_n$-fixed behavior can be interpreted as sensitivity analysis of $T_{BC}(\hat{\lambda}_n)$ when $\hat{\lambda}_n$ is different from λ_0 (Duan, 1993). By $\hat{\lambda}_n$-fixed we mean ignoring the randomness of $\hat{\lambda}_n$ but not the effect of changing parameter values and sample size. This is handled by writing $\hat{\lambda}_n = \lambda_0 + \Delta\tau(\hat{\lambda}_n)$, where Δ is the standardized $\hat{\lambda}_n$ and $\tau(\hat{\lambda}_n)$ is the standard

Table 4.1: Simulated $sd's$ of the ith element (t_i) of the Box-Cox it T-ratio

Model 1, $\lambda_0 = 0.0$, $sd(T_0) = 1.0426$						Model 2, $\lambda_0 = 0.05$, $sd(T_{0i}) = 1.8974$				
$\sigma_0 \ddots \mu_0$	0.1	1.0	5.0	20.0		β_0	σ_0	t_1	t_2	t_3
0.1	1.0633	1.0565	1.0605	–		21,11,1	0.01	1.9481	1.9318	1.9608
1.0	1.0570	1.0579	1.0608	1.0585			0.1	1.9575	1.9293	1.9334
10.0	1.0593	1.0617	1.0529	1.0563			1.0	1.9543	1.8920	1.9403
							5.0	1.9631	1.8981	1.9393
Model 3: $\lambda_0 = 0.0$, first row is $sd(T_0)$						2^3 factorial design: $\lambda_0 = 0.0$, first row is $sd(T_0)$				
σ_0	t_1	t_2	t_3	t_4		σ_0	t_1	t_2	t_3	t_4
	1.1619	0.4230	1.4230	1.4230			1.8371	2.4646	2.4054	2.4054
0.001	1.1655	1.4501	1.4259	1.4365		0.001	1.9054	2.5975	2.5657	2.5849
1.0	1.1604	1.4262	1.4285	1.4249		0.1	1.8636	2.4878	2.4771	2.4568
1.0	1.1927	1.4304	1.4234	1.4116		1.0	1.8683	2.4624	2.4357	2.4341
5.0	1.1975	1.4136	1.4172	1.4156		5.0	1.8448	2.4252	2.4290	2.4740

deviation of $\hat{\lambda}_n$. Fixing $\hat{\lambda}_n$ means fixing Δ but not $\tau(\hat{\lambda}_n)$ with respect to η, σ_0 and n. This is practically meaningful as, for example, when $n = 25$ one obtains $\hat{\lambda}_n = 0.5$, which can be used to transform the current data set or future data set of the same size and from the same situations. However, when data set is doubled in size or the experimental setting is changed, one definitely would not use the same '0.5' to transform the data, instead would reestimate the transformation value.

5.1 Theoretical Results

Theorem 5.1. *Assume that the conditions of Corollary 2.1 are satisfied, then*

$$E[T_{BC}(\hat{\lambda}_n)|\Delta \text{ fixed}] = E(T_0) + O_p(n^{-1}), \tag{5.1}$$

$$Var[T_{BC}(\hat{\lambda}_n)|\Delta \text{ fixed}] = Var(T_0) + 2\Gamma(\Delta) + O_p(n^{-1}), \tag{5.2}$$

where $\Gamma(\Delta) = \Delta\tau(\hat{\lambda}_n)E(T_0U_1') - \Delta^2\tau^2(\hat{\lambda}_n)E(T_0U_2')$.

Proof. This is straightforward following Corollary 2.1.

When θ_0 is small, U_1 and U_2 in Corollary 2.1 can be approximated by $\tilde{U}_1$ and $\tilde{U}_2$ of Theorem 4.1. Thus, $E(T_0U_1')$ and $E(T_0U_2')$ can be easily approximated, which gives,

$$\Gamma(\Delta) \approx n\lambda_0^{-1}\Delta\tau(\hat{\lambda}_n)(X'X)^{-1}X'DX(X'X)^{-1} - \frac{1}{2}\sigma_0^{-2}\Delta^2\tau^2(\hat{\lambda}_n)\|Qa\|^2(X'X)^{-1}, \tag{5.3}$$

where $D = \text{Diag}\{\phi_i - \bar{\phi}\}_{n\times n}$, and $a = \lambda_0^{-2}(1 + \lambda_0\eta)\log(1 + \lambda_0\eta)$.

Letting $\lambda_0 \to 0$ in (5.2) gives an expression for corresponding to a log-transformation. Using (5.2) in connection with Theorem 3.1, we can summarize the behavior of $\Gamma(\Delta)$ for Δ fixed as follows.

(i) For structured models with small σ_0, $\Gamma(\Delta)$ becomes negative, suggesting that the $\hat{\lambda}_n$-fixed variance of $T_{BC}(\hat{\lambda}_n)$ is smaller than $Var(T_0)$. That is, a misspecified transformation deflates the variance of Box-Cox T-ratio. This can be seen more clearly from the limit

$$\lim_{\sigma_0 \to 0} \Gamma(\Delta) = -\frac{1}{2}\Delta^2 (X'X)^{-1}.$$

However, as it is very likely that Δ takes values in the interval (-2, 2), that is, $\hat{\lambda}_n$ is within two standard deviations of λ_0, this deflation factor will be small, especially when n is large.

(ii) For unstructured models, the second term in (5.2) vanishes, and when σ_0 is small, the ith diagonal element of $\Gamma(\Delta)$ is positive if $\Delta(\phi_i - \bar{\phi}) > 0$ and negative if $\Delta(\phi_i - \bar{\phi}) < 0$, suggesting that $\hat{\lambda}_n$-fixed variance of ith element of $T_{BC}(\hat{\lambda}_n)$ is accordingly larger or smaller than the ith diagonal element of $Var(\lambda_0)$. This point becomes more evident from the following limit,

$$\lim_{\lambda_0 \to 0} \Gamma(\Delta) = \frac{1}{\sqrt{2}} n\Delta \|\phi - \bar{\phi} 1_n\|^{-1} (X'X)^{-1} X'DX (X'X)^{-1}.$$

Hence when n is small the $\hat{\lambda}_n$-fixed effect may not be negligible for an unstructured model with σ_0 small relative to the spread in means. Further, it is easy to see that $tr\Gamma(\Delta) = 0$ or close to 0, suggesting that the total variance $trVar[T_{BC}(\hat{\lambda}_n)]$ does not depend much on the value of $\hat{\lambda}_n$.

(iii) For a one-mean model, $\Gamma(\Delta) = 0$, indicating that when n is small the $\hat{\lambda}_n$-fixed effect of a estimating transformation will be very small in this case.

(iv) Finally when σ_0 is large, all models behave like a one-mean model and the effect of transform-ation misspecification is very small as evidenced by $\lim_{\sigma_0 \to \infty} \Gamma(\Delta) = 0$ for $\lambda_0 = 0$. This limit suggests that for other transformations $\Gamma(\Delta)$ is small when σ_0 is large but θ_0 is small.

Combining the results of Section 3 and 5, we conclude that in the cases that $\hat{\lambda}_n$ behaves poorly such as one-mean models with small σ_0, $T_{BC}(\hat{\lambda}_n)$ is very robust to the changes of $\hat{\lambda}_n$, while in the cases that $T_{BC}(\hat{\lambda}_n)$ is sensitive to the changes in $\hat{\lambda}_n$ such as unstructured models with σ_0 small relative to the mean-spread, $\hat{\lambda}_n$ behaves very well. This is an important conclusion; it sheds light on the validity of Box-Cox methodology. The practical implication of these results is profound: the Box-Cox transformation methodology performs well in most of the statistical inferential situations.

5.2 Monte Carlo Simulations

The three models in Section 3 are used again with similar parameter settings. The simulation results reported in Tables 5.1 are the simulated $\hat{\lambda}_n$-fixed standard deviations of $T_{BC}(\hat{\lambda}_n)$ or the

Table 5.1: Simulated $\hat{\lambda}_n$-fixed sd of $T_BC(\hat{\lambda}_n)$

Model 1: $\lambda_0 = 0.0$, $sd(T_0) = 1.0342$						Model 2: $\lambda_0 = 0.01, \beta_0 = (12, 11, 10)'$ $sd(T_{0i}) = 1.8974$				
Δ	σ_0	$\mu_0 = 0.1$	$\mu_0 = 1.0$	$\mu_0 = 10$		Δ	σ_0	t_1	t_2	t_3
3	0.1	1.0585	1.0598	1.0416		3	0.01	2.5872	1.6783	1.0891
-3		1.0580	1.0657	1.0494		-3		1.0967	1.6608	2.5874
3	1.0	1.0492	1.0493	1.0493		3.0	0.1	2.5739	1.6925	1.0837
-3		1.0632	1.0461	1.0612		-3		1.1153	1.6704	2.6194
3	10.	1.0485	1.0520	1.0627		3	2.0	2.2163	1.8336	1.5580
-3		1.0460	1.0463	1.0396		-3		1.5304	1.8360	2.2273
Model 3: $\lambda_0 = 0.0, \Delta = \pm 2, sd(T_0) = (1.1619, 1.4230, 1.4230, 1.4230)$										
σ_0	t_1	t_2	t_3	t_4		σ_0	t_1	t_2	t_3	t_4
0.001	1.0719	1.2924	1.2823	1.3016		1.0	1.1635	1.4044	1.3778	1.4059
	1.0637	1.2983	1.2849	1.3096			1.1460	1.3921	1.3871	1.3759
0.1	1.0498	1.2934	1.2969	1.2819		10.0	1.1447	1.3846	1.3846	1.3883
	1.0685	1.2878	1.2899	1.2940			1.1630	1.3919	1.3992	1.3704

element of $T_{BC}(\hat{\lambda}_n)$ indicated by t_i in the table when $\hat{\lambda}_n = \lambda_0 + \Delta\tau(\hat{\lambda}_n)$ with $\Delta = \pm 2, \pm 3$. The value of $\tau(\hat{\lambda}_n)$ is calculated using (3.2) or (3.4). More extensive simulation results are available from the author.

The results for a one-mean model show that $sd[T_{BC}(\hat{\lambda}_n)|\hat{\lambda}_n \text{fixed}]$ can be well approximate by $sd(T_0)$ for any situations, irrespective to the size of the difference $\hat{\lambda}_n - \lambda_0$. This means that for the one-mean model $T_{BC}(\hat{\lambda}_n)$ is very robust against transformation misspecification. The results for the three-means model show that the $\hat{\lambda}_n$-fixed effect on the individual variance can be significant but not on the total variance, which is consistent with the theory. In model 3, our theory suggests that the $\hat{\lambda}_n$-fixed sd of $T_{BC}(\hat{\lambda}_n)$ be smaller than $sd(T_0)$ when σ_0 is small. Simulations show that it is indeed smaller, but only slightly if $|\Delta| \cdot$ 2. The effect is small when σ_0 is moderate to large. The $\hat{\lambda}_n$-fixed sd of $T_{BC}(\hat{\lambda}_n)$ can also be easily approximated by (5.1). Calculations (not reported) show that it is very accurate.

Appeared in: Test, 1999, 8, 167-190.

APPENDIX: Proof of Theorem 2.1

A second-order Taylor expansion of $\bar{\boldsymbol{\Psi}}_1(y, \hat{\psi}_n)$ around $(\beta_0, \lambda_0, \sigma_0)$ gives

$$\begin{aligned} 0 &= +\bar{\boldsymbol{\Psi}}_1 + \dot{\boldsymbol{\Psi}}_{11}(\hat{\beta}_n\beta_0) + \dot{\boldsymbol{\Psi}}_{12}(\hat{\lambda}_n - \lambda_0) + \dot{\boldsymbol{\Psi}}_{13}(\hat{\sigma}_n - \sigma_0) + \tfrac{1}{2}[I_p \quad (\hat{\beta}_n - \beta_0)']\ddot{\boldsymbol{\Psi}}_{111}(\hat{\beta}_n - \beta_0) \\ &+\tfrac{1}{2}\ddot{\Psi}_{122}(\hat{\lambda}_n - \lambda_0)^2 + \tfrac{1}{2}\ddot{\Psi}_{133}(\hat{\sigma}_n - \sigma_0)^2 + \ddot{\Psi}_{112}(\hat{\beta}_n - \beta_0)(\hat{\lambda}_n - \lambda_0) \\ &+\ddot{\boldsymbol{\Psi}}_{113}(\hat{\beta}_n - \beta_0)(\hat{\sigma}_n - \sigma_0) + \ddot{\boldsymbol{\Psi}}_{123}(\hat{\lambda}_n - \lambda_0)(\hat{\sigma}_n - \sigma_0) + O_p(n^{-3/2}). \end{aligned} \tag{A1}$$

First-order Taylor expansions of $\bar{\boldsymbol{\Psi}}_1(y, \hat{\psi}_n)$ and $\bar{\boldsymbol{\Psi}}_3(y, \hat{\psi}_n)$ around $(\beta_0, \lambda_0, \sigma_0)$ yield

$$\hat{\beta}_n - \beta_0 = -\mathbf{A}_{11}^{-1}[\bar{\boldsymbol{\Psi}}_1 + \mathbf{A}_{11}(\hat{\lambda}_n - \lambda_0)] + O_p(n^{-1}). \tag{A2}$$

$$\hat{\sigma}_n - \sigma_0 = -\mathbf{A}_{33}^{-1}[\bar{\boldsymbol{\Psi}}_3 + \mathbf{A}_{32}(\hat{\lambda}_n - \lambda_0)] + O_p(n^{-1}). \tag{A3}$$

Substituting A2 and A3 into A1 for the terms of order $O_p(n^{-1})$ and replacing $\ddot{\boldsymbol{\Psi}}_{ijk}$ by $\mathbf{B_{ijk}}$ give

$$\begin{aligned} \hat{\beta}_n - \beta_0 = \quad & -\mathbf{A}_{11}^{-1}\{\bar{\boldsymbol{\Psi}}_1 - (\dot{\boldsymbol{\Psi}}_{11} - \mathbf{A}_{11})\mathbf{A}_{11}^{-1}\bar{\boldsymbol{\Psi}}_1 - \dot{\boldsymbol{\Psi}}_{13}\mathbf{A}_{33}^{-1}\bar{\boldsymbol{\Psi}}_3 + \tfrac{1}{2}[I_p \quad (\mathbf{A}_{11}^{-1}\bar{\boldsymbol{\Psi}}_1)]'\mathbf{B}_{111}\mathbf{A}_{11}^{-1}\bar{\boldsymbol{\Psi}}_1 \\ & +\mathbf{\tfrac{1}{2}B_{133}(A_{33}^{-1}\bar{\Psi}_3)^2 + B_{113}A_{11}^{-1}A_{33}^{-1}\bar{\Psi}_1\bar{\Psi}_3\} - A_{11}^{-1}\{\dot{\Psi}_{12} - (\dot{\Psi}_{11} - A_{11})A_{11}^{-1}A_{12}} \\ & -\dot{\boldsymbol{\Psi}}_{13}\mathbf{A}_{33}^{-1}\mathbf{A}_{32} + [I_p \quad (\mathbf{A}_{11}^{-1}\bar{\boldsymbol{\Psi}}_1)]'\mathbf{B}_{111}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} + \mathbf{B}_{133}\mathbf{A}_{33}^{-2}\mathbf{A}_{32}\bar{\boldsymbol{\Psi}}_3 - \mathbf{B}_{112}\mathbf{A}_{11}^{-1}\bar{\boldsymbol{\Psi}}_1 \\ & +\mathbf{B}_{113}\mathbf{A}_{11}^{-1}\mathbf{A}_{33}^{1}(\bar{\boldsymbol{\Psi}}_1\mathbf{A}_{32} + \mathbf{A}_{12}\bar{\boldsymbol{\Psi}}_3) - \mathbf{B}_{123}\mathbf{A}_{33}^{-1}\bar{\boldsymbol{\Psi}}_3\}(\hat{\lambda}_n - \lambda_0) - \mathbf{A}_{11}^{-1}\{\tfrac{1}{2}\mathbf{B}_{122} \\ & +\tfrac{1}{2}[I_p \quad (\mathbf{A}_{11}^{-1}\mathbf{A}_{12})]'\mathbf{B}_{111}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} + \tfrac{1}{2}\mathbf{B}_{133}(\mathbf{A}_{33}^{-1}\mathbf{A}_{32})^2 - \mathbf{B_{112}A_{11}^{-1}A_{12}} \\ & +\mathbf{B}_{113}\mathbf{A}_{11}^{-1}\mathbf{A}_{33}^{-1}\mathbf{A}_{12}\mathbf{A}_{32} - \mathbf{B}_{123}\mathbf{A}_{33}^{-1}\mathbf{A}_{32}\}(\hat{\lambda}_n - \lambda_0)^2 + O_p(n^{-3/2}). \end{aligned} \tag{A4}$$

Now, taking expectation of A4 treating $\hat{\lambda}$ as fixed gives expansions of $\beta_u(\hat{\lambda}_n)$ and $\hat{\beta}_n - \beta_u(\hat{\lambda}_n)$. Letting $\hat{\lambda}_n = \lambda_0$ in the expansion of $\beta_u(\hat{\lambda}_n)$ and $\hat{\beta}_n - \beta_u(\hat{\lambda}_n)$ results in an expansion for $\hat{\beta}_{n0} - \beta_u(\lambda_0)$. Thus,

$$\hat{\beta}_n - \beta_u(\hat{\lambda}_n) = \hat{\beta}_{n0} + \text{the second-order term} + O_p(n^{-3/2}). \tag{A5}$$

Now, considering $\hat{\sigma}_n^{-1}$ as a function of $\hat{\sigma}_n^2$, a first-order Taylor expansion gives

$$\hat{\sigma}_n^{-1} = \hat{\sigma}_{n0}^{-1} - \tfrac{1}{2}\sigma_0^{-3}(\hat{\sigma}_n^2 - \hat{\sigma}_{n0}^2) + O_p(n^{-1}), \tag{A6}$$

and considering $\hat{\sigma}_n^2$ as a function of $\hat{\lambda}_n$, we have by a second-order Taylor expansion around λ_0

$$\hat{\sigma}_n^2 = \hat{\sigma}_{n0}^2 + \dot{r}(\lambda_0)(\hat{\lambda}_n - \lambda_0) + \ddot{r}(\lambda_0)(\hat{\lambda}_n - \lambda_0)^2, \tag{A7}$$

where $\dot{r}(\lambda_0)$ and $\ddot{r}(\lambda_0)$ are the first and second derivatives of $\hat{\sigma}_n^2$ with respect to $\hat{\lambda}_n$ evaluated at λ_0. Substituting (A7) into (A6) and multiplying the resulted expression by (A5) side by side and then by $\sqrt{n}$ gives the first part of Theorem 2.1. The second part is obtained by substituting (A2) and (A3) into

$$0 = \bar{\boldsymbol{\Psi}}(y, \hat{\boldsymbol{\Psi}}_n) = \bar{\boldsymbol{\Psi}}_2 + \dot{\boldsymbol{\Psi}}_{21}(\hat{\beta}_n - \beta_0) + \dot{\boldsymbol{\Psi}}_{22}(\hat{\lambda}_n - \lambda_0) + \dot{\boldsymbol{\Psi}}_{23}(\hat{\sigma}_n - \sigma_0) + O_p(n^{-1}).$$

The final expressions of U_1 and U_2 are given as follows

$$\begin{aligned} U_1 = \; & \tfrac{1}{2}\sqrt{n}\sigma_0^{-3}\mathbf{A}_{11}^{-1}\bar{\mathbf{\Psi}}_1\dot{r}(\lambda_0) - \sqrt{n}\sigma_0^{-1}\mathbf{A}_{11}^{-1}\{\dot{\mathbf{\Psi}}_{12} - \mathbf{A}_{12}(\dot{\mathbf{\Psi}}_{11} - \mathbf{A}_{11})\mathbf{A}_{11}{-}1\mathbf{A}_{12} - \dot{\mathbf{\Psi}}_{13}\mathbf{A}_{33}^{-1}\mathbf{A}_{32} \\ & +[I_p \quad (\mathbf{A}_{11}^{-1}\bar{\mathbf{\Psi}}_1)]'\mathbf{B}_{111}\mathbf{A}_{11}{-}1\mathbf{A}_{12} + \mathbf{B}_{133}\mathbf{A}_{33}{-}2\mathbf{A}_{32}\bar{\mathbf{\Psi}}_3 - \mathbf{B}_{112}\mathbf{A}_{11}^{-1}\bar{\mathbf{\Psi}}_1 \\ & \mathbf{B}_{113}\mathbf{A}_{11}^{-1}\mathbf{A}_{33}{-}1(\bar{\mathbf{\Psi}}_1\mathbf{A}_{32} + \mathbf{A}_{12}\bar{\mathbf{\Psi}}_3) - \mathbf{B}_{123}\mathbf{A}_{33}^{-1}\bar{\mathbf{\Psi}}_3\}, \\ U_2 = \; & \tfrac{1}{4}\sqrt{n}\sigma_0^{-3}\mathbf{A}_{11}^{-1}\bar{\mathbf{\Psi}}_1\ddot{r}(\lambda_0). \end{aligned}$$

ACKNOKWLEDGEMENTS

The author would like to thank the referee for the constructive and insightful comments that lead to significant improvements on the manuscript.

REFERENCES

Atkinson, A. C. (1985). *Plots, Transformations and Regression*, Oxford Univ. Press.

Bickel, P. J. (1984). Comment on "The Analysis of Transformed Data," *J. Amer. Statist. Assoc.*, 79, 315-316.

Bickel, P. J., and Doksum, K. A. (1981). An analysis of transformations revisited, *J. Amer. Statist. Assoc.*, 76, 296-311.

BOX, G. E. P., and COX, D. R. (1964). An analysis of transformations (with discussion), *J. Roy. Statist. Soc. Ser. B,* 26, 211-252.

BOX, G. E. P., and COX, D. R. (1982). An analysis of transformations revisited, rebutted, *J. Amer. Statist. Assoc.*, 77, 209-210.

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

Cohen, A, and Sackrowitz, H. B. (1987). An approach to inference following model selection with applications to transformation-based and adaptive inference, *J. Amer. Statist. Assoc.*, 82, 1123-1130.

Cox, D. R., and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion), *J. Roy. Statist. Soc. Ser. B*, 49, 1-39.

Draper, N. R., and Cox, D. R. (1969). On distribution and their transformation to normality, *J. Roy. Statist. Soc. Ser. B*, 31, 472-476.

Duan, N. (1993). Sensitivity analysis for Box-Cox power transformation model: contrast parameters, *Biometrika*, 80, 885-897.

Hinkley, D. V. (1975). On power transformations to symmetry, *Biometrika*, 62, 101-111.

Hinkley, D. V., and Runger, G. (1984). The analysis of transformed data (with discussion), *J. Amer. Statist. Assoc.*, 79, 302-320.

Hooper, P. M., and Yang, Z. (1997). Confidence intervals following Box-Cox transformation, *Canadian Journal of Statistics*, 25, 401-416.

Lawrance, A. J. (1987). A note on the variance of the Box-Cox regression transformation estimate, *Applied Statistics*, 36, 221-223.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review, *The Statistician*, 41, 169-178.

Yang, Z. (1992). *Inference following Box-Cox transformation*, Ph.D. Dissertation, Department of Statistics and Applied Probability, University of Alberta, Canada.

Yang, Z. (1996). Some asymptotic results on Box-Cox transformation methodology, *Commun. Statist.-Theory and Meth.*, 25, 403-414.

Yang, Z. (1997). More on the estimation of Box-Cox transformation, *Commun. Statist.- Simula.*, 26, 1063-1074.

Yang, Z. (1998). An alternative approximation to the variance of transformation score, *J. Statist. Comput. Simul.*, to appear.