Appeared in: Canadian Journal of Statistics, 2008, 36, 301-319.

The Canadian Journal of Statistics
Vol. ?, No. ?, 2007, Pages ???-???
La revue canadienne de statistique

Inference for general parametric functions in Box-Cox-type transformation models

Key words and phrases: Box-Cox transformation; confidence interval; marginal effects; percentile function; robustness; survivor function; tests; variance inflation.
MSC 2000: Primary 62J05; secondary 62F25, 62E20.

Abstract: The authors consider a simple but general method of inference for a parametric function of the Box-Cox-type transformation model, which is built upon the normal inference theories when the transformation parameter is known and takes the estimation of this parameter into account. This method quickly leads to test statistics and confidence intervals for i) a linear combination of the regression coefficients, ii) a linear combination of the scaled regression coefficients, iii) marginal effects on the median of an original response, iv) a general percentile function of an original response, and v) the survivor function. They show, through Monte Carlo simulations, that the finite sample performance of the new method is often superior to the commonly used delta method, and that the new method is robust to mild departures from normality of error distributions. They illustrate the new method with a numerical example.

Résumé : The authors consider a simple but general method of inference for a parametric function of the Box-Cox-type transformation model, which is built upon the normal inference theories when the transformation parameter is known and takes the estimation of this parameter into account. This method quickly leads to test statistics and confidence intervals for i) a linear combination of the regression coefficients, ii) a linear combination of the scaled regression coefficients, iii) marginal effects on the median of an original response, iv) a general percentile function of an original response, and v) the survivor function. They show, through Monte Carlo simulations, that the finite sample performance of the new method is often superior to the commonly used delta method, and that the new method is robust to mild departures from normality of error distributions. They illustrate the new method with a numerical example.

1. INTRODUCTION

The standard Box-Cox transformation model (Box & Cox, 1964) takes the form

$$h(\mathbf{Y}, \lambda) = \mathbf{X}\beta + \varepsilon, \tag{1}$$

where $h(\mathbf{Y}, \lambda)$ is an $n \times 1$ vector of transformed responses $h(Y_i, \lambda)$ with h being a general monotonic function, λ denotes the transformation parameter(s), $\mathbf{X}$ is an $n \times k$ matrix containing the values of k regressors, β is a $k \times 1$ vector of parameters, and ε is an $n \times 1$ vector of independent and identically distributed (iid) normal random variables with mean zero and variance σ^2.

There are many transformation functions available in the literature. The most popular one is the Box-Cox power transformation which takes the form

$$h(y, \lambda) = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0, \\ \log y, & \lambda = 0. \end{cases} \tag{2}$$

However, this transformation is incompatible with the normality assumption due to the well-known truncation problem, unless the transformation parameter λ equals zero. If exact normality is crucial, such as in the proofs of certain theorems, one must seek alternative transformations. A closely related transformation, called the *dual power transformation*, that also works for positive variables is given in Yang (2006) which takes the form

$$h(y, \lambda) = \begin{cases} (y^\lambda - y^{-\lambda})/(2\lambda), & \lambda \neq 0, \\ \log y, & \lambda = 0. \end{cases} \tag{3}$$

This transformation is shown to possess properties similar to those of the Box-Cox power transformation, but does not suffer from the truncation problem.

When observations can be both positive and negative, neither the Box-Cox power transformation nor the dual power transformation is appropriate. In this case, several alternative transformations are available. See, for example, Manly (1976), John & Draper (1980), Bickel & Doksum (1981), Burbidge, Magee & Robb (1988), MacKinnon & Magee (1990), and Yeo & Johnson (2000).

Since the seminal work of Box & Cox (1964), there have been many developments (including debates) and applications of the Box-Cox transformation methodology, in a wide variety of fields including, other than statistics, economics, engineering, medicine and biological sciences. However, a simple but general and reliable method of inference for the Box-Cox transformation model seems still lacking. In this case, one typically relies on the delta method for inference which is known to possibly suffer from a poor finite sample performance.

In this paper, we develop a general inference method for model (1) that works for all transformations with which the normality assumption is possible. In particular, we develop general methods for testing hypotheses about, and constructing confidence intervals for, $\psi = g(\lambda, \beta, \sigma^2)$, a general smooth function of the model parameters. In some important special cases, we also give an alternative method, aiming for better finite sample properties. The key factors that may contribute to the improved finite sample performance are: a) the alternative method is able to take advantage of the exact λ-known distribution of the underlying statistic if it exists, and b) the alternative method is able to alleviate the effect of nonlinearity of the function g.

As applications of the general theory, we develop tests and confidence intervals for i) a linear combination of the regression coefficients, ii) a linear combination of the scaled regression coefficients, iii) marginal effects on the median of an original response, iv) a general percentile function of an original response, and v) the survivor function. All results are simple and computationally they involve little more than the calculations of the full MLEs of the model parameters.

The reliability (or finite sample performance) of the new methods is assessed through Monte Carlo simulation, based on two transformation functions: the Box-Cox power transformation given in (2) and the dual power transformation given in (3), with errors being normal, truncated normal, or normal-gamma mixture. We find that the finite sample performance of the new method is often superior to the commonly used delta method, in particular in the cases of (iv) and (v), and is robust against mild departures from normality of error distributions.

The rest of the paper is organized as follows. Section 2 presents the general results. Section 3 presents five applications. Section 4 presents Monte Carlo results for the finite sample performance of the new methods. Section 5 provides a numerical illustration. Section 6 concludes the paper.

2. THE GENERAL METHODS

The log likelihood function in relation to the original observations $\mathbf{Y}$ is, ignoring the constant,

$$\ell(\lambda, \beta, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}[h(\mathbf{Y}, \lambda) - \mathbf{X}\beta)]'[h(\mathbf{Y}, \lambda) - \mathbf{X}\beta)] + \sum_{i=1}^{n} \log h_y(Y_i, \lambda), \tag{4}$$

where $h_y(Y_i, \lambda) = \partial h(Y_i, \lambda)/\partial Y_i$. Maximization of the log likelihood is usually carried out in two steps. First, for a given λ, $\ell(\lambda, \beta, \sigma^2)$ is maximized by

Appeared in: Canadian Journal of Statistics, 2008, 36, 301-319.

$$\hat{\beta}(\lambda) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'h(\mathbf{Y}, \lambda) \text{ and } \hat{\sigma}^2(\lambda) = n^{-1}\|\mathbf{M}h(\mathbf{Y}, \lambda)\|^2, \tag{5}$$

where $\|\cdot\|$ is the Euclidean norm and $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ with $\mathbf{I}_n$ being the $n \times n$ identity matrix. Then, substituting $\hat{\beta}(\lambda)$ and $\hat{\sigma}^2(\lambda)$ into (4) gives the concentrated log likelihood for λ as,

$$\ell_c(\lambda) = -\frac{n}{2}\log(\|\mathbf{M}h(\mathbf{Y}, \lambda)\|^2) + \sum_{i=1}^{n}\log h_y(Y_i, \lambda).$$

The unconstrained MLE $\hat{\lambda}$ of λ thus maximizes $\ell_c(\lambda)$. Finally, substituting $\hat{\lambda}$ back into $\hat{\beta}(\lambda)$ and $\hat{\sigma}^2(\lambda)$ for λ gives the MLEs of β and λ denoted as $\hat{\beta}(\hat{\lambda})$ and $\hat{\sigma}^2(\hat{\lambda})$.

Suppose in general our inference concerns $\psi = g(\lambda, \beta, \sigma^2)$, where g is a general smooth function of the model parameters. Examples of such functions include linear combinations of regression coefficients β, linear combinations of the scaled regression coefficients β/σ, percentile functions of a response y_0 at regressor values x_0, marginal effects on the median of a response y_0, survivor functions, etc. We denote by g_β and g_{σ^2} the partial derivatives of g with respect to β and σ^2, respectively

When λ is known, β and σ^2 are estimated, respectively, by the constrained MLEs $\hat{\beta}(\lambda)$ and $\hat{\sigma}^2(\lambda)$ given in (5), which gives the constrained MLE for ψ as $\hat{\psi}(\lambda) \equiv g(\lambda, \hat{\beta}(\lambda), \hat{\sigma}^2(\lambda))$. If λ is unknown and is estimated by its MLE $\hat{\lambda}$, then β and σ^2 are estimated, respectively, by $\hat{\beta}(\hat{\lambda})$ and $\hat{\sigma}^2(\hat{\lambda})$, which gives an unconstrained MLE of ψ as $\hat{\psi}(\hat{\lambda}) \equiv g(\hat{\lambda}, \hat{\beta}(\hat{\lambda}), \hat{\sigma}^2(\hat{\lambda}))$. Denote the derivative of $\hat{\psi}(\lambda)$ with respect to λ by $\hat{\psi}_\lambda(\lambda)$. We now present our general theory. The proofs of the theorems are sketched in the Appendix.

THEOREM 1. *Assume i) the function g is differentiable in all its three arguments; ii) $\hat{\psi}_\lambda(\lambda)$ is bounded in probability, and iii) consistency and asymptotic normality hold for the MLEs of Model (1) parameters. We have*

$$\sqrt{n}[\hat{\psi}(\lambda) - \psi] \xrightarrow{D} N(0, v^2), \text{ and}$$

$$\sqrt{n}[\hat{\psi}(\hat{\lambda}) - \psi] \xrightarrow{D} N(0, v^2 + \kappa^2\tau^2),$$

where $v^2 = \lim_{n\to\infty}[n\sigma^2 g'_\beta(\mathbf{X}'\mathbf{X})^{-1}g_\beta] + 2\sigma^4 g^2_{\sigma^2}$, $\kappa = \lim_{n\to\infty} \mathrm{E}[\hat{\psi}_\lambda(\lambda)]$, and τ^2 is the asymptotic variance of $\sqrt{n}(\hat{\lambda} - \lambda)$.

Theorem 1 says that estimation of λ inflates the variance of the estimator for ψ by a factor $\kappa^2\tau^2$, where the derivative $\hat{\psi}_\lambda(\lambda)$ plays a key role in the magnitude of this variance inflation. The situations where this derivative is small, so that the variance inflation is small, are of particular interest (see Chen, Lockhart & Stephens, 2002). The results of Theorem 1 rely on the consistency and asymptotic normality of the MLEs of the model parameters, which are not addressed in this paper as the focus of the paper is on the inference for a parametric function. Interested readers may refer to Hernandez & Johnson (1980), Bickel & Doksum (1981), Carroll (1982), Carroll & Ruppert (1984), Foster, Tian & Wei (2001) and Chen, Lockhart & Stephens (2002) for similar large sample frameworks in which the estimates of model parameters are consistent and asymptotically normal.

Theorem 1 suggests two methods of inference for ψ. **Method 1** follows directly from the second part of the theorem, i.e., inference for ψ proceeds with

$$T_1(\hat{\lambda}, \psi) = \frac{\sqrt{n}[\hat{\psi}(\hat{\lambda}) - \psi]}{\sqrt{\hat{v}^2 + \hat{\kappa}^2\hat{\tau}^2}}, \tag{6}$$

where the estimates $\hat{v}^2$ and $\hat{\kappa}^2$ are obtained by evaluating the finite sample expressions for v^2 and κ^2 (dropping the limits) at the full MLEs of all the parameters. The variance estimate $\hat{\tau}^2$ is obtained according to the method described later. An asymptotic two-sided level α test for testing

$H_0 : \psi = \psi_0$ rejects the null hypothesis when $|T_1(\hat{\lambda}, \psi_0)| > Z_{\alpha/2}$, and an asymptotic $100(1-\alpha)\%$ confidence interval (CI) for ψ is

$$\hat{\psi}(\hat{\lambda}) \pm Z_{\alpha/2}\sqrt{(\hat{v}^2 + \hat{\kappa}^2\hat{\tau}^2)/n}, \tag{7}$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

Though the inference method described above is easier to apply than the commonly used delta method (described in the Appendix), its finite sample performance may not be better, as, similar to the delta method, it also works directly on the function g, i.e., estimating g (by plugging-in the parameter estimates) and finding the asymptotic variance of the estimate. Hence, Method 1 may suffer from the same problem of poor finite sample performance as does the delta method.

In many applications in engineering, medicine, biological sciences, etc., where only a small data set is available, it is highly desirable to have inference procedures with good small sample properties. One way to do this is to take advantage of the exact finite sample distribution (if it exists) of the λ-known statistic

$$T_0(\lambda, \psi) = \frac{\sqrt{n}[\hat{\psi}(\lambda) - \psi]}{\hat{v}(\lambda)},$$

where $\hat{v}(\lambda)$ is the MLE of v at the known value of λ. The idea is to find simple modifications of $T_0(\hat{\lambda}, \psi)$ so that the modified statistic has a good match in distribution with $T_0(\lambda, \psi)$. We call this **Method 2**, which is developed by the following arguments. Denote the mean, variance and the distribution of $T_0(\lambda, \psi)$ by μ_T, σ_T^2 and F_T, respectively. Following the proof of Theorem 1, we have

$$T_0(\hat{\lambda}, \psi) = T_0(\lambda, \psi) + \frac{\kappa}{v}\sqrt{n}(\hat{\lambda} - \lambda) + o_p(1),$$

which leads to $\mathrm{E}[T_0(\hat{\lambda}, \psi)] = \mu_T + o(1)$ and $\mathrm{Var}[T_0(\hat{\lambda}, \psi)] = \sigma_T^2 + v^{-2}\kappa^2\tau^2 + o(1)$. Thus, the two statistics match in means (with an error of o(1)) but not in variances. In order to be able to refer to the finite sample distribution when making inference about ψ, it is necessary to modify $T_0(\hat{\lambda}, \psi)$ so that its first two moments match those of $T_0(\lambda, \psi)$. As normally $\mu_T \neq 0$, a change in variance causes a shift in mean. The modified statistic thus takes the form

$$T_2(\hat{\lambda}, \psi) = \frac{T_0(\hat{\lambda}, \psi) - \mu_T(1 - \hat{v}_f)}{\hat{v}_f}, \tag{8}$$

where $\hat{v}_f = \{1 + (\frac{\hat{\kappa}\,\hat{\tau}}{\hat{v}\sigma_T})^2\}^{1/2}$. Thus, inference for ψ can be based on $T_2(\hat{\lambda}, \psi)$ which refers to the distribution F_T instead of $N(0,1)$ in Method 1. An approximate two-sided level α test of $H_0 : \psi = \psi_0$ rejects the null hypothesis when $T_2(\hat{\lambda}, \psi) < F_T^{1-\alpha/2}$ or $T_2(\hat{\lambda}, \psi) > F_T^{\alpha/2}$, and an approximate $100(1-\alpha)\%$ CI for ψ is

$$\left\{\hat{\psi}(\hat{\lambda}) - \frac{\hat{v}}{\sqrt{n}}\left(\mu_T(1 - \hat{v}_f) + \hat{v}_f F_T^{\alpha/2}\right),\ \hat{\psi}(\hat{\lambda}) - \frac{\hat{v}}{\sqrt{n}}\left(\mu_T(1 - \hat{v}_f) + \hat{v}_f F_T^{1-\alpha/2}\right)\right\}, \tag{9}$$

where $F_T^{1-\alpha/2}$ and $F_T^{\alpha/2}$ are the lower and upper $100(\alpha/2)\%$ points of F_T.

A simple method for computing $\hat{\tau}^2$. By a theorem of Patefield (1977), the inverse concentrated information matrix is equal to the covariance matrix obtained by inverting the full information matrix and taking the appropriate submatrix. A simple estimate of τ^2 is given by

$$\hat{\tau}^2 = -n\left(\frac{\partial^2}{\partial\lambda^2}\ell_c(\hat{\lambda})\right)^{-1}, \tag{10}$$

which can be handled numerically by many statistical software. Carroll & Ruppert (1988, p129) recommended that the above second derivative can be approximated by

$$\frac{\partial^2}{\partial\lambda^2}\ell_c(\hat{\lambda}) \approx \frac{1}{\epsilon^2}\left[\ell_c(\hat{\lambda} + \epsilon) + \ell_c(\hat{\lambda} - \epsilon) - 2\ell_c(\hat{\lambda})\right],$$

where ϵ is a small positive number, say $\epsilon = 0.01$.

This simple way of estimating the variance τ^2 makes the application of Theorem 1 more convenient. Once the full MLEs of model parameters are available, the rest are just simple substitutions. The same type of inference using other methods usually is much more computationally involved. For example, the delta method involves the calculation of the full information matrix, and the likelihood ratio method involves a nonlinear maximization subject to a nonlinear constraint.

Special cases. There are interesting special cases of the g function that merit further study. We formally treat one case here and more can be seen from the applications. The idea here is to make use of the special form of the function g to reduce the effect of nonlinearity of it, thus improving the finite sample performance of the inference procedures. Consider $g(\lambda, \beta, \sigma) = g(\lambda, c'\xi)$ where c is a vector of constants and $\xi = \{\beta', \sigma\}'$. If g, now considered as a function of $c'\xi$ only, is invertible with

$$c'\xi = g^{-1}(\lambda, \psi) \equiv f(\lambda, \psi)$$

then inference for ψ is made much easier and perhaps better in its finite sample performance (see the applications in the next section and the discussions therein). This is because the asymptotic distribution of the statistic is much less dependent on the nonlinearity of the function g in β and σ^2. Specifically, the two derivatives: g_β and g_{σ^2}, disappear as seen in the following theorem.

THEOREM 2. *Assume the conditions of Theorem 1 hold. Assume further i) $g(\lambda, \beta, \sigma^2) = g(\lambda, c'\xi)$, ii) $c'\xi = g^{-1}(\lambda, \psi) \equiv f(\lambda, \psi)$ exists, iii) $f_\lambda(\lambda, \psi) = (\partial/\partial\lambda) f(\lambda, \psi)$ exists, and iv) $\hat{\beta}_\lambda(\lambda)$ and $\hat{\sigma}_\lambda(\lambda)$ are bounded in probability. Then, we have,*

$$\sqrt{n}[c'\hat{\xi}(\lambda) - f(\lambda, \psi)] \xrightarrow{D} N(0, v_0^2), \text{ and}$$

$$\sqrt{n}[c'\hat{\xi}(\hat{\lambda}) - f(\hat{\lambda}, \psi)] \xrightarrow{D} N(0, v_0^2 + \kappa_0^2\tau^2), \tag{11}$$

where $v_0^2 = \lim_{n\to\infty}[n\sigma^2 c_1'(\mathbf{X}'\mathbf{X})^{-1}c_1 + \frac{1}{2}c_2^2\sigma^2]$ *with* $(c_1', c_2) = c'$, $\kappa_0 = \lim_{n\to\infty} \mathrm{E}[c'\hat{\xi}_\lambda(\lambda)] - f_\lambda(\lambda, \psi)$, *and* τ^2 *is given in Theorem 1.*

Again, Theorem 2 shows that there is a variance inflation in the estimation of ψ caused by the estimation of the transformation parameter. From the result (11) of Theorem 2, it is clear that inferences for ψ (Method 1) can be carried out based on

$$T_1(\hat{\lambda}, \psi) = \frac{\sqrt{n}[c'\hat{\xi}(\hat{\lambda}) - f(\hat{\lambda}, \psi)]}{\sqrt{\hat{v}_0^2 + \hat{\kappa}_0^2\hat{\tau}^2}},$$

where $\hat{v}_0^2 = n\hat{\sigma}^2(\hat{\lambda})c_1'(\mathbf{X}'\mathbf{X})^{-1}c_1 + \frac{1}{2}c_2^2\hat{\sigma}^2(\hat{\lambda})$ and $\hat{\kappa}_0 = c'\hat{\xi}_\lambda(\hat{\lambda}) - f_\lambda(\hat{\lambda}, \hat{\psi})$. A two-sided level α test for testing H_0: $\psi = \psi_0$ rejects the null hypothesis when $|T_1(\hat{\lambda}, \psi_0)| > Z_{\alpha/2}$, and an asymptotic $100(1-\alpha)\%$ CI for ψ is

$$\left\{g[\hat{\lambda}, L(\hat{\lambda})], \quad g[\hat{\lambda}, U(\hat{\lambda})]\right\}$$

where $L(\hat{\lambda}) = c'\hat{\xi}(\hat{\lambda}) - Z_{\alpha/2}\{(\hat{v}_0^2 + \hat{\kappa}_0^2\hat{\tau}^2)/n\}^{\frac{1}{2}}$ and $U(\hat{\lambda}) = c'\hat{\xi}(\hat{\lambda}) + Z_{\alpha/2}\{(\hat{v}_0^2 + \hat{\kappa}_0^2\hat{\tau}^2)/n\}^{\frac{1}{2}}$.

Parallel to the developments that lead to the statistic (8) and the CI (9), if the finite sample distribution F_T of the λ-known statistic

$$T_0(\lambda, \psi) = \frac{\sqrt{n}[c'\hat{\xi}(\lambda) - f(\lambda, \psi)]}{\hat{\sigma}(\lambda)\sqrt{nc_1'(\mathbf{X}'\mathbf{X})^{-1}c_1 + \frac{1}{2}c_2^2}},$$

is completely known, with mean μ_T and variance σ_T^2, one can easily develop a similar method (Method 2) for this case by modifying $T_0(\hat{\lambda}, \psi)$. The modified statistic is

$$T_2(\hat{\lambda}, \psi) = \frac{T_0(\hat{\lambda}, \psi) - \mu_T(1 - \hat{v}_{f0})}{\hat{v}_{f0}}, \tag{12}$$

where $\hat{v}_{f0} = \{1 + (\frac{\hat{\kappa}_0 \hat{\tau}}{\hat{v}_0 \sigma_T})^2\}^{1/2}$. The modified CI for ψ is

$$\left\{ g[\hat{\lambda}, L_0(\hat{\lambda})], \quad g[\hat{\lambda}, U_0(\hat{\lambda})] \right\}, \tag{13}$$

where $L_0(\hat{\lambda}) = c'\hat{\xi}(\hat{\lambda}) - \frac{\hat{v}_0}{\sqrt{n}}[\mu_T(1 - \hat{v}_{f0}) + \hat{v}_{f0} F_T^{\alpha/2}]$, and $U_0(\hat{\lambda}) = c'\hat{\xi}(\hat{\lambda}) - \frac{\hat{v}_0}{\sqrt{n}}[\mu_T(1 - \hat{v}_{f0}) + \hat{v}_{f0} F_T^{1-\alpha/2}]$.

3. APPLICATIONS OF THE GENERAL METHODS

We now present some interesting applications of the general inference methodology for the Box-Cox transformation model presented above. We note that τ^2 is the asymptotic variance of $\sqrt{n}(\hat{\lambda} - \lambda)$ where $\hat{\lambda}$ is the full MLE of λ, and that $\hat{\tau}^2$ is the estimate of τ^2 given by (10). Hence, these quantities take the same form no matter which theorem is applied and what application is being considered. We also note that, in order to apply Theorems 1 or 2, it is only necessary to derive finite sample expressions for v^2 and κ^2 for a given g so that their MLEs can be obtained.

3.1 Regression coefficients

There has been some debate on which regression coefficients should be of inferential focus after a transformation has been applied to the response. Bickel & Doksum (1981) argued inference should be made on β with λ being treated as another genuine parameter like β and σ. Box & Cox (1982), and Hinkley & Runger (1984) argued that λ is just a scale of data analysis and once it is chosen, the subsequent analysis should be carried out and the effects be interpreted based on this scale. Thus, the regression coefficients concerned should be $\beta(\hat{\lambda})$ defined and interpreted based on the chosen scale $\hat{\lambda}$. For example, if $\hat{\lambda}$ is found to be 0.5, then one should fit the model $Y(0.5) = \mathbf{X}\beta(0.5) + \sigma(0.5)e$ and interpret all the effects based on $\lambda = 0.5$, i.e., the square-root transformation. Results reported in the literature show that if inferences concern β, the usual normal-theory inference methods are not valid. However, if inferences concern $\beta(\hat{\lambda})$, the usual inference methods remain asymptotically valid. See, for example, Bickel & Doksum (1981), Hinkley & Runger (1984), Hooper & Yang (1997) and Yang (1999).

We now apply the theorems presented in the earlier section and outline arguments supporting the asymptotic validity of the usual inference method when inference concerns $\beta(\hat{\lambda})$, and present a simple way for correcting the usual method when inference concerns β. Consider a general linear function $\psi = a'\beta$, for a fixed vector a. First, when λ is known, we have the following statistic

$$T_0(\lambda, \psi) = \frac{a'\hat{\beta}(\lambda) - a'\beta}{\hat{\sigma}(\lambda)\sqrt{a'(\mathbf{X}'\mathbf{X})^{-1}a}},$$

which is exactly $\sqrt{\frac{n}{n-k}}\, t_{n-k}$, where t_{n-k} denotes a t random variable with $n-k$ degrees of freedom. Tests and confidence intervals for $a'\beta$ can easily be constructed. An exact $100(1-\alpha)\%$ CI for $a'\beta$ is: $a'\hat{\beta}(\lambda) \pm t_{n-k}^{\alpha/2}\{na'(\mathbf{X}'\mathbf{X})^{-1}a/(n-k)\}^{\frac{1}{2}}\hat{\sigma}(\lambda)$, where $t_{n-k}^{\alpha/2}$ denotes the upper $\alpha/2$ quantile of t_{n-k}.

When λ is unknown and is estimated by $\hat{\lambda}$, Method 1 of Theorem 1 leads to the asymptotic $N(0,1)$ statistic: $T_1(\hat{\lambda}, \psi) = \sqrt{n}[a'\hat{\beta}(\hat{\lambda}) - a'\beta]/\sqrt{\hat{v}^2 + \hat{\kappa}^2\hat{\tau}^2}$, where $\hat{v}^2 = n\hat{\sigma}^2(\hat{\lambda})a'(\mathbf{X}'\mathbf{X})^{-1}a$ and $\hat{\kappa} = a'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'h_\lambda(\mathbf{Y}; \hat{\lambda})$; an approximate two-sided level α test of $H_0: \psi = \psi_0$ that rejects the null hypothesis when $|T_1(\hat{\lambda}, \psi_0)| > Z_{\alpha/2}$; and an approximate $100(1-\alpha)\%$ CI for $a'\beta$ as:

$$a'\hat{\beta}(\hat{\lambda}) \pm Z_{\alpha/2}\{(\hat{v}^2 + \hat{\kappa}^2\hat{\tau}^2)/n\}^{\frac{1}{2}}.$$

Similarly, Method 2 of Theorem 1 leads to the following statistic $T_2(\hat{\lambda}, \psi) = T_0(\hat{\lambda}, \psi)/\hat{v}_f$, where $\hat{v}_f = \{1 + (\frac{\hat{\kappa}\hat{\tau}}{\hat{v}\sigma_T})^2\}^{1/2}$ and $\sigma_T^2 = \frac{n}{n-k-2}$; an approximate two-sided level α test of $H_0: \psi = \psi_0$ that rejects the null hypothesis when $|T_2(\hat{\lambda}, \psi_0)| > (\frac{n}{n-k})^{1/2} t_{n-k}^{\alpha/2}$; and an approximate $100(1-\alpha)\%$ CI for $a'\beta$ as:

$$a'\hat{\beta}(\hat{\lambda}) \pm t_{n-k}^{\alpha/2} \left\{(\hat{v}^2\sigma_T^2 + \hat{\kappa}^2\hat{\tau}^2)/(\sigma_T^2(n-k))\right\}^{\frac{1}{2}}.$$

Finally, if inference concerns $a'\beta(\hat{\lambda})$, where $\beta(\hat{\lambda}) = E[\hat{\beta}(\hat{\lambda})|\hat{\lambda}]$ as in Hinkley & Runger (1984), then it can be easily proved, along the line of proof for Theorem 1, that the following statistic

$$T^*(\hat{\lambda}, \psi) = \frac{a'\hat{\beta}(\hat{\lambda}) - a'\beta(\hat{\lambda})}{\hat{\sigma}(\hat{\lambda})\sqrt{a'(\mathbf{X}'\mathbf{X})^{-1}a}},$$

has an asymptotic $N(0,1)$ distribution, provided that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{h_\lambda(\mathbf{Y},\lambda) - \mathrm{E}[h_\lambda(\mathbf{Y},\lambda)|\hat{\lambda}]\} \xrightarrow{p} 0$. A similar result holds for the definition of $\beta(\hat{\lambda})$ given in Hooper & Yang (1997) (see also Yang, 1999). Thus, it follows that when inference concerns regression parameters that are defined on the scale determined by the chosen transformation $\hat{\lambda}$, the usual inference is asymptotically valid.

3.2 Scaled regression coefficients

Continuing the debate (discussed in Section 3.1) over the parameter of interests following a response transformation, Chen, Lockhart & Stephens (2002) argue that inference should be carried out on the scaled regression coefficients. They provide some large sample arguments showing that these scaled regression coefficients are much more stable with respect to the change in the transformation scale than the original regression coefficients. Theorem 1 of this paper quickly leads to some similar results as those given by Chen, Lockhart & Stephens (2002). Suppose now inferences concern the scaled regression coefficients: $\psi = g(\lambda, \beta, \sigma^2) = a'\beta/\sigma$. We have, $g_\beta = a\sigma^{-1}$, $g_{\sigma^2} = -\frac{1}{2}a'\beta\sigma^{-3}$ and $\hat{\psi}(\lambda) = a'\hat{\beta}(\lambda)/\hat{\sigma}(\lambda)$, which give, $v^2 \approx na'(\mathbf{X}'\mathbf{X})^{-1}a + \frac{1}{2\sigma^2}(a'\beta)^2$, and $\hat{\psi}_\lambda(\lambda) = a'\hat{\beta}_\lambda(\lambda)\hat{\sigma}^{-1}(\lambda) - a'\hat{\beta}(\lambda)\hat{\sigma}_\lambda(\lambda)\hat{\sigma}^{-2}(\lambda)$, where $\hat{\beta}_\lambda(\lambda) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'h_\lambda(\mathbf{Y},\lambda)$ and $\hat{\sigma}_\lambda(\lambda) = \frac{1}{n\hat{\sigma}(\lambda)}h'(\mathbf{Y},\lambda)\mathbf{M}h_\lambda(\mathbf{Y},\lambda)$. With the above expressions, Theorem 1 can be easily implemented for making inference about $a'\beta/\sigma$. In particular, the following asymptotic $N(0,1)$ pivotal quantity can be used,

$$T_1(\hat{\lambda}, \psi) = \frac{\sqrt{n}[a'(\hat{\beta}(\hat{\lambda})/\hat{\sigma}(\hat{\lambda})) - a'(\beta/\sigma)]}{\sqrt{\hat{v}^2 + \hat{\kappa}^2\hat{\tau}^2}},$$

where $\hat{v}^2$ is v^2 given above evaluated at the full MLEs of β and σ^2, and $\hat{\kappa}^2 = \hat{\psi}_\lambda(\hat{\lambda})$. It is interesting to further investigate the key quantity in the variance inflation factor, $\hat{\psi}_\lambda(\hat{\lambda})$. Simulation results given in Yang (2002a) show that this variance inflation is small. CIs for ψ can be constructed as in (7) and (9).

3.3 Marginal effects

In usual linear regressions, β itself represents the marginal effect of X on $\mathrm{E}(Y)$. In the Box-Cox-type regression model, however, β no longer has such a meaning. A similar measure is the marginal effect on the median of the response Y. We have, $\mathrm{Med}[h(Y,\lambda)] = x'\beta$, which, following monotonicity of h in Y, gives $h(\mathrm{Med}(Y),\lambda) = x'\beta$, and $\mathrm{Med}(Y) = h^{-1}(x'\beta,\lambda) \equiv f(x'\beta,\lambda)$. The f notation, here, is the same as that used before. The marginal effect of an individual explanatory variable x_j on Y can be generally expressed as follows

$$\psi = g(\lambda, \beta, \sigma^2) = a'\partial f(x'\beta, \lambda)/\partial x = a'\beta f_\eta(\eta, \lambda),$$

where $\eta = x'\beta$, and a is now a vector of elements with 1 in the jth position and zeros otherwise. This gives $\hat{\psi}(\lambda) = a'\hat{\beta}(\lambda)f_\eta(\hat{\eta}(\lambda),\lambda)$, $\hat{\psi}(\hat{\lambda}) = a'\hat{\beta}(\hat{\lambda})f_\eta(\hat{\eta}(\hat{\lambda}),\hat{\lambda})$, and $\hat{\psi}_\lambda(\lambda) = a'\hat{\beta}_\lambda(\lambda)f_\eta(\hat{\eta}(\lambda),\lambda) +$

$a'\hat{\beta}(\lambda)[f_{\eta\eta}(\hat{\eta}(\lambda),\lambda)x'\hat{\beta}_\lambda(\lambda)+f_{\eta\lambda}(\hat{\eta}(\lambda),\lambda)]$, where $f_{\eta\lambda}(\hat{\eta}(\lambda),\lambda) = \partial^2 f(\eta,\lambda)/\partial\eta\partial\lambda|_{\eta=\hat{\eta}(\lambda)}$. Now, $g_\beta = af_\eta(x'\beta,\lambda) + a'\beta f_{\eta\eta}(x'\beta,\lambda)x$, and $g_{\sigma^2} = 0$ as g is free of σ^2. We have $v^2 \approx n\sigma^2 g'_\beta(\mathbf{X}'\mathbf{X})^{-1}g_\beta$.

Thus, with all the expressions given above, Theorem 1 leads to the following asymptotic $N(0,1)$ statistic for ψ,

$$T_1(\hat{\lambda},\psi) = \frac{\sqrt{n}[a'\hat{\beta}(\hat{\lambda})f_\eta(\hat{\eta}(\hat{\lambda}),\hat{\lambda}) - \psi]}{\sqrt{\hat{v} + \hat{\kappa}^2\hat{\tau}^2}}.$$

Tests and confidence intervals for ψ can be constructed in a similar way as in the above applications. The case of known λ is simple and hence is not discussed.

3.4 Percentile function

From Sections 3.1 to 3.3, we have seen some applications of Theorem 1. We now present a case where Theorem 2 can be applied, from which the importance of the results in Theorem 2 can be seen, in particular, the alternative procedure. Consider inferences for ψ: the 100pth percentile of a future observation y_0 at x_0. As h is a monotonic function, it follows that $h(\psi,\lambda) = x_0'\beta+\sigma z_p$, where z_p is the 100pth percentile of $N(0,1)$ (note the difference between z_p and $Z_{\alpha/2}$ introduced earlier). One immediately notices that this case falls into the framework of Theorem 2 with $c_1 = x_0$ and $c_2 = z_p$. Thus, $\sqrt{n}[x_0'\hat{\beta}(\hat{\lambda})+\hat{\sigma}(\hat{\lambda})z_p - h(\psi,\hat{\lambda})] \xrightarrow{D} N(0, v_0^2+\kappa_0^2\tau^2)$, where $v_0^2 \approx n\sigma^2 x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + \frac{1}{2}z_p^2\sigma^2$ and $\kappa_0 \approx x_0'\hat{\beta}_\lambda(\lambda) + \hat{\sigma}_\lambda(\lambda)z_p - h_\lambda(\psi,\lambda)$. The statistic for inference for ψ becomes,

$$T_1(\hat{\lambda},\psi) = \frac{\sqrt{n}[x_0'\hat{\beta}(\hat{\lambda}) + \hat{\sigma}(\hat{\lambda})z_p - h(\psi,\hat{\lambda})]}{\sqrt{\hat{v}_0^2 + \hat{\kappa}_0^2\hat{\tau}^2}}.$$

An asymptotic $100(1-\alpha)\%$ CI for ψ is

$$\left\{h^{-1}(L(\hat{\lambda}),\hat{\lambda}), \quad h^{-1}(U(\hat{\lambda}),\hat{\lambda})\right\},$$

where $L(\hat{\lambda})$ and $U(\hat{\lambda})$ take on the minus and plus parts of $x_0'\hat{\beta}(\hat{\lambda})+\hat{\sigma}(\hat{\lambda})z_p \pm Z_{\alpha/2}\{(\hat{v}^2+\hat{\kappa}^2\hat{\tau}^2)/n\}^{\frac{1}{2}}$, respectively. Inferences for the case of λ-known proceed in a similar fashion by dropping the variance inflation factor and replacing everywhere $\hat{\lambda}$ by λ.

We now apply the results of (12) and (13) to provide a possibly improved inference procedure. Letting $a_0 = [x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0]^{1/2}$, $c_0 = (na_0^2 + \frac{1}{2}z_p^2)^{1/2}$, we have when λ is given,

$$T_0(\lambda,\psi) = \frac{\sqrt{n}[x_0'\hat{\beta}(\lambda) + \hat{\sigma}(\lambda)z_p - h(\psi,\lambda)]}{\hat{\sigma}(\lambda)c_0}.$$

It is easy to show that $T_0(\lambda,\psi) \sim \frac{a_0 n}{c_0\sqrt{n-k}}t_{n-k}(-z_p/a_0) + \frac{\sqrt{n}z_p}{c_0}$, where $t_{n-k}(\delta)$ denotes a noncentral t random variable with noncentrality parameter δ and '$\sim$' denotes 'is distributed as'. Thus, the finite sample distribution F_T of $T_0(\lambda,\psi)$ is completed specified, and the results of (12) and (13) are applicable. From the first two moments of a noncentral t distribution, we obtain

$$\begin{aligned}\mu_T &= \frac{\sqrt{n}z_p}{c_0}\left(1 - \frac{\sqrt{n}\Gamma[(n-k-1)/2]}{\sqrt{2}\Gamma[(n-k)/2]}\right), \quad \text{and}\\ \sigma_T^2 &= \frac{n^2(a_0^2+z_p^2)}{(n-k-2)c_0^2} - \frac{n^2 z_p^2}{2c_0^2}\left(\frac{\Gamma[(n-k-1)/2]}{\Gamma[(n-k)/2]}\right)^2,\end{aligned}$$

which give the variance factor $\hat{v}_{f0}$, and the modified statistic $T_2(\hat{\lambda},\psi) = [T_0(\hat{\lambda},\psi) - \mu_T(1-\hat{v}_{f0})]/\hat{v}_{f0}$. Thus, the modified $100(1-\alpha)\%$ CI for ψ is

$$\left\{g[\hat{\lambda}, L_0(\hat{\lambda})], \quad g[\hat{\lambda}, U_0(\hat{\lambda})]\right\}, \tag{14}$$

with all the necessary quantities specified as follows

$$
\begin{aligned}
L_0(\hat\lambda) &= x_0'\hat\beta(\hat\lambda) + \hat\sigma(\hat\lambda)z_p - \tfrac{\hat v_0}{\sqrt{n}}[\mu_T(1-\hat v_{f0}) + \hat v_{f0}F_T^{\alpha/2}],\\
U_0(\hat\lambda) &= x_0'\hat\beta(\hat\lambda) + \hat\sigma(\hat\lambda)z_p - \tfrac{\hat v_0}{\sqrt{n}}[\mu_T(1-\hat v_{f0}) + \hat v_{f0}F_T^{1-\alpha/2}],\\
F_T^{1-\alpha/2} &= \tfrac{a_0 n}{c_0\sqrt{n-k}}t_{n-k}^{1-\alpha/2}(-z_p/a_0) + \tfrac{\sqrt{n}z_p}{c_0}, \text{ and } F_T^{\alpha/2} = \tfrac{a_0 n}{c_0\sqrt{n-k}}t_{n-k}^{\alpha/2}(-z_p/a_0) + \tfrac{\sqrt{n}z_p}{c_0}.
\end{aligned}
$$

We note that the CIs given above are very easy to compute as compared with the other approaches such as the likelihood ratio and delta method. More interestingly, the CI given by (14) has a better finite sample performance as shown in the following section. See the discussions in the introductory section and the end of the next subsection for theoretical reasons why this is so. When $p = 0.5$, $z_p = 0$ and the results reduces to those for the median. Yang (2002b) provides an alternative result for the case which works specifically for the Box-Cox power transformation. We also note that confidence intervals for percentile curves is an important topic in the context of medical studies where they are often referred to as reference curves.

3.5 Survivor function

The survivor function for a response Y_0 (say) with the corresponding regressor values x_0 is the probability for Y_0 to exceed a given value y_0. Let $z_0(\lambda,\beta,\sigma) = [h(y_0,\lambda) - x_0'\beta]/\sigma$. We have the parameter of interest, i.e., the survivor function,

$$\psi = g(\lambda,\beta,\sigma^2) = 1 - \Phi[z_0(\lambda,\beta,\sigma)],$$

where Φ denotes the cumulative distribution function of the standard normal random variable. Let ϕ be the probability density function corresponding to Φ. We have, $g_\beta = \frac{1}{\sigma}\phi[z_0(\lambda,\beta,\sigma)]x_0$ and $g_{\sigma^2} = \frac{1}{2\sigma^2}z_0(\lambda,\beta,\sigma)\phi[z_0(\lambda,\beta,\sigma)]$, which gives, $v^2 \approx \phi[z_0(\lambda,\beta,\sigma)]^2\left(nx_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + \frac{1}{2}z_0^2(\lambda,\beta,\sigma)\right)$. Furthermore, let $\hat z_0(\lambda) = z_0[\lambda,\hat\beta(\lambda),\hat\sigma(\lambda)]$, then $\hat\psi(\lambda) = 1 - \Phi[\hat z_0(\lambda)]$, which gives $\hat\psi_\lambda(\lambda) = -\phi[\hat z_0(\lambda)]\hat z_{0\lambda}(\lambda)$, where $\hat z_{0\lambda}(\lambda) = \frac{1}{\hat\sigma(\lambda)}[h_\lambda(y_0,\lambda) - x_0'\hat\beta_\lambda(\lambda) - \hat z_0(\lambda)\hat\sigma_\lambda(\lambda)]$. Evaluating v^2 and $\hat\psi_\lambda(\lambda)$ at the full MLEs of model parameters leads to an asymptotic $N(0,1)$ statistic as that given by (6), which in turn gives a confidence interval as that given by (7).

Alternatively, a simpler and perhaps better method can be obtained by starting with inferences for $\Phi^{-1}(1-\psi)$, where ψ is the survivor function defined above and Φ^{-1} denotes the inverse of the Φ function. Theorem 1 leads immediately to the following asymptotic $N(0,1)$ pivotal quantity

$$T(\hat\lambda,\psi) = \frac{\sqrt{n}[\hat z_0(\hat\lambda) - \Phi^{-1}(1-\psi)]}{\sqrt{v^2+\kappa^2\tau^2}}$$

where $v^2 = \lim_{n\to\infty}[nx_0'(\mathbf{X}'\mathbf{X})^{-1}x_0] + \frac{1}{2}z_0^2(\lambda,\beta,\sigma)$, and $\kappa = \lim_{n\to\infty}E[\hat z_{0\lambda}(\lambda)]$. Thus, an asymptotic $100(1-\alpha)\%$ CI for ψ is given by

$$\{1-\Phi[U(\hat\lambda)],\quad 1-\Phi[L(\hat\lambda)]\}, \tag{15}$$

where $L(\hat\lambda) = \hat z_0(\hat\lambda) - Z_{\alpha/2}\sqrt{(\hat v^2+\hat\kappa^2\hat\tau^2)/n}$, $U(\hat\lambda) = \hat z_0(\hat\lambda) + Z_{\alpha/2}\sqrt{(\hat v^2+\hat\kappa^2\hat\tau^2)/n}$, $\hat v^2 = nx_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + \frac{1}{2}\hat z_0^2(\hat\lambda)$, and $\hat\kappa = \hat z_{0\lambda}(\hat\lambda)$.

Although this result does not exactly follow Theorem 2, it falls into the spirit of discussions regarding the effect of nonlinearity of the function g. Thus, it is expected that the CI (15) perform well. This point is confirmed by the simulation results given in the next section. It is well known that the performance of the large-sample normal-theory method can be improved by considering a proper transformation. For example, if ψ is a quantity that must be positive (e.g., a quantile duration or event time), then it is better to assume $\log\hat\psi$, rather than $\hat\psi$, follows an asymptotic normal distribution; if ψ is restricted to be between 0 and 1 (e.g., a survivor function), confidence intervals based on $\log[\hat{\psi}/(1-\hat{\psi})]$ usually perform better than those based on $\hat{\psi}$. The survivor function takes values between 0 and 1. The delta method works directly on this function, as does our first method. Thus, the second method should have a better finite sample performance. See Hahn & Meeker (1991, p239) for more discussions.

Another important point as illustrated by the case of the percentile function discussed in Section 3.4 is whether the method is able to take advantage of the exact distribution of the λ-known statistic. If it does, the finite sample performance of the λ-unknown inference procedures can be improved.

4. MONTE CARLO SIMULATION

In this section, we report some results of a Monte Carlo experiment to i) show the finite sample properties of the confidence intervals described in Section 3, ii) compare the confidence intervals based on our method with those based on the delta method, and iii) investigate the truncation effects or more generally the robustness of the methods against departures from normality of the error distribution. The data is generated from the following model

$$h(Y_i, \lambda) = \beta_0 + \beta_1 X_i + \varepsilon_i, \; i = 1, \cdots, n,$$

where ε_i are iid with mean zero and variance σ^2, h is either the Box-Cox power transformation or the dual power transformation, and the X_i values are generated uniformly from the interval [0,5].

The simulation process is as follows. First, we generate a sample of n observations from the $N(0, \sigma^2)$ or some other non-normal population with mean zero and variance σ^2, then convert these observations into Y_i through the above model relationship, then estimate the model and calculate the confidence interval. Repeating this process 10,000 times, the proportion of the intervals in 10,000 that contain ψ gives a simulated coverage probability of the confidence interval.

One point to observe is that all the transformation functions mentioned in the introduction are compatible with the exact normality assumption made on ε_i, except the Box-Cox power transformation where ε_i must be bounded from below (left truncated) if $\lambda > 0$ or bounded from above (right truncated) if $\lambda < 0$ to guarantee the positivity of Y_i. The truncation effect for the Box-Cox power transformation is negligible when this bound is large which occurs when σ is small, or model means are large, or λ is small. Nevertheless, this raises an interesting question: what happens to our methods if truncation is necessary, or more generally ε_i are not exactly normal. It is well known that if the errors ε_i are not normal, but still iid with zero mean, and finite four-plus moments, the parameter estimates based on the normal likelihood can still be consistent (see White, 1994), giving the so-called quasi-MLEs. Thus, the effect of truncation or non-normality is on the standard error estimates, and hence on the inferences for the model parameters.

We investigate these issues by generating ε_i/σ values from (i) the standard normal distribution, (ii) a (standardized) truncated normal distribution with 10% truncation on the left tail or on the right tail, or 5% truncation on each side, and (iii) a normal-gamma mixture with 90% of ε_i/σ from $N(0, 1)$ and the remaining 10% from a (standardized) gamma distribution with both parameters equal to one. The following parameter values are chosen: $(\beta_0, \beta_1) = (3.5, 1)$ if $\lambda > 0$ and $(-3.5, -1)$ if $\lambda < 0$, $\lambda = (.5, .25, 0, -.25, -.5)$, $\sigma = (.01, .1, .5)$, $n = (25, 50, 100)$, and $p = (.05, .25, .5, .75, .95)$. When the dual power transformation is used, the cases of negative λ are dropped due to symmetry. For the cases of percentile and survivor functions, the value for x_0 is set to $(1, 2.5)'$.

Selected simulation results are presented in Figures 1-3, where `New1` refers to CI based on our Method 1, `New2` refers to CI based on our Method 2, and `Delta` refers to CI based on the delta method. More complete results are available from the authors upon request. In each plot of each figure, the vertical scale is the simulated coverage probability and the horizontal scale is the **index** of parameter configurations which will be explained for each case below. For the effects of truncation or non-normality in general, we concentrate on the cases of percentile and survivor functions as these two cases provide the most interesting and important applications of our results.

Appeared in: Canadian Journal of Statistics, 2008, 36, 301-319.

CIs for $\psi = a'\beta$, $\psi = c'\beta/\sigma$, and the marginal effect. To conserve space, we only report in Figure 1 the results corresponding to the Box-Cox power transformation with normal errors, where the horizontal index $(1-15)$ represents the 15 possible combinations of the λ and σ values arranged by first increasing the λ value for a given σ value, and then increasing the σ value. The plots in the first row of Figure 1 summarize the coverage probabilities of the CIs for $a'\beta$, based on the new methods and the delta method. The results show that the method referring to the standard normal distributions (`New1`) gives a CI that is comparable with the CI based on the delta method. However, the new method referring to t distribution (`New2`) has a better finite sample performance. The coverage probabilities do not change much with the parameter values (except when n is small), and quickly converge to their nominal levels when n increases. The plots in the second row of Figure 1 summarize the coverage probabilities of the CIs for a linear combination of the scaled regression coefficients, and the plots in the third row of Figure 1 give the coverage probabilities for the CIs for the marginal effects. The results for the latter two cases show that our Method 1 gives confidence intervals that possess comparable finite sample properties as those given by the delta method. However, the new method is easier to implement.

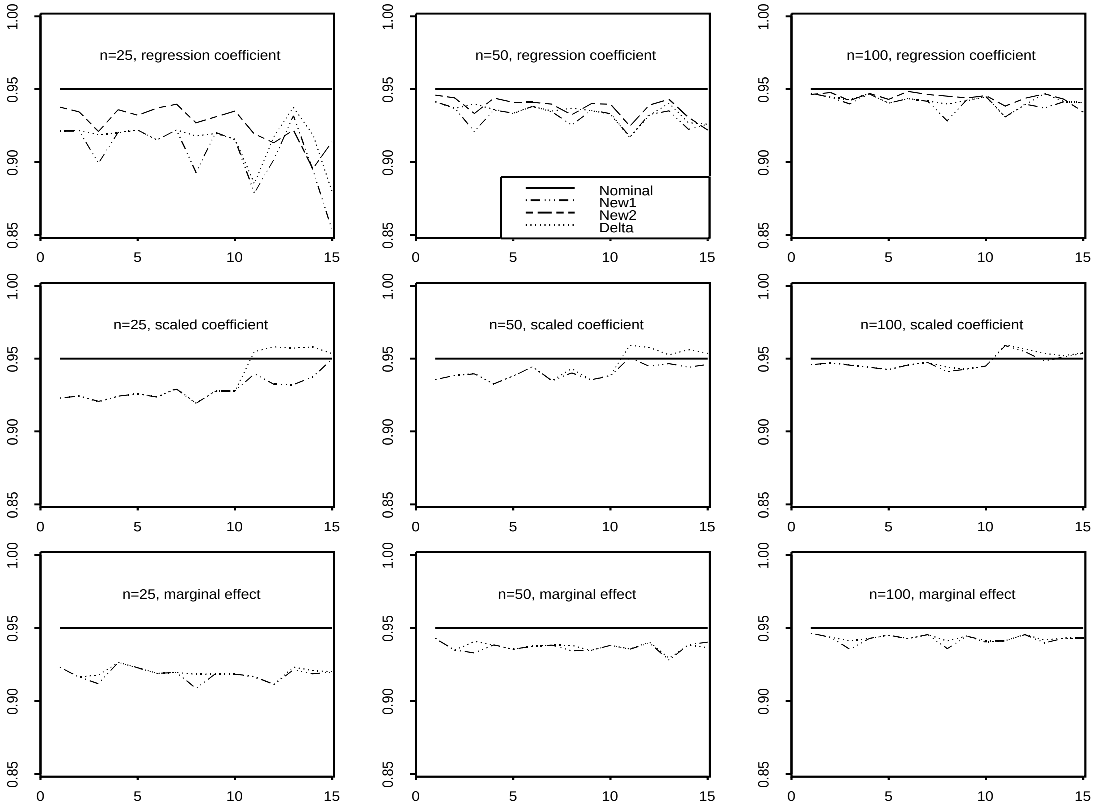

Figure 1: Coverage Probabilities of 95% CIs for Regression Coefficients, Scaled Regression Coefficients, and Marginal Effects, under Normal Errors

CI for percentile function. The CIs developed in Section 3.4 provide the most interesting applications of Theorem 2. However, it is not yet clear whether the reality conforms with the theoretical prediction. The Monte Carlo simulation results given in Figure 2 ((a)-(e)) for Box-Cox power transformation and Figure 3 ((a)-(d)) for dual power transformation confirm the theoretical prediction. In Figure 2, the horizontal index $(1-75)$ represents the 75 combinations of the values for λ, σ, and p (the probability value corresponding to the percentile considered), arranged with λ increasing first, followed by σ, and then by p, whereas in Figure 3 it reduces to $1-45$ as the two negative values for λ are dropped.

In the case of normal errors, the results show that our Method 2 based on Theorem 2 has led to a CI with much better small sample performance than the CIs based on Method 1 and the delta method. This is especially true when p is either small or large (two ends of the plots). In these cases, the difference in coverage probabilities of the CIs can be substantial when n is not large, with those from Method 2 much closer to their nominal level 0.95. In all situations, the coverage probability converges to its nominal level when sample size increases. In the case of truncated normal errors, the results (not fully reported for brevity) show that left truncation results in a higher coverage for a low percentile and a lower coverage for a high percentile. Right truncation results in the opposite patterns, and a symmetric truncation does not affect the coverage probability much. Very interestingly, in all cases increasing sample size seems to improve the coverage. In the case of a normal-gamma mixture, the results (not fully reported for brevity) show that the effects of non-normality is more severe for the delta method and our Method 1 than for our Method 2. The effect is very small for our Method 2 except for the upper extreme percentiles. In general, the results show that our method is quite robust against mild departures from normality of error distributions.

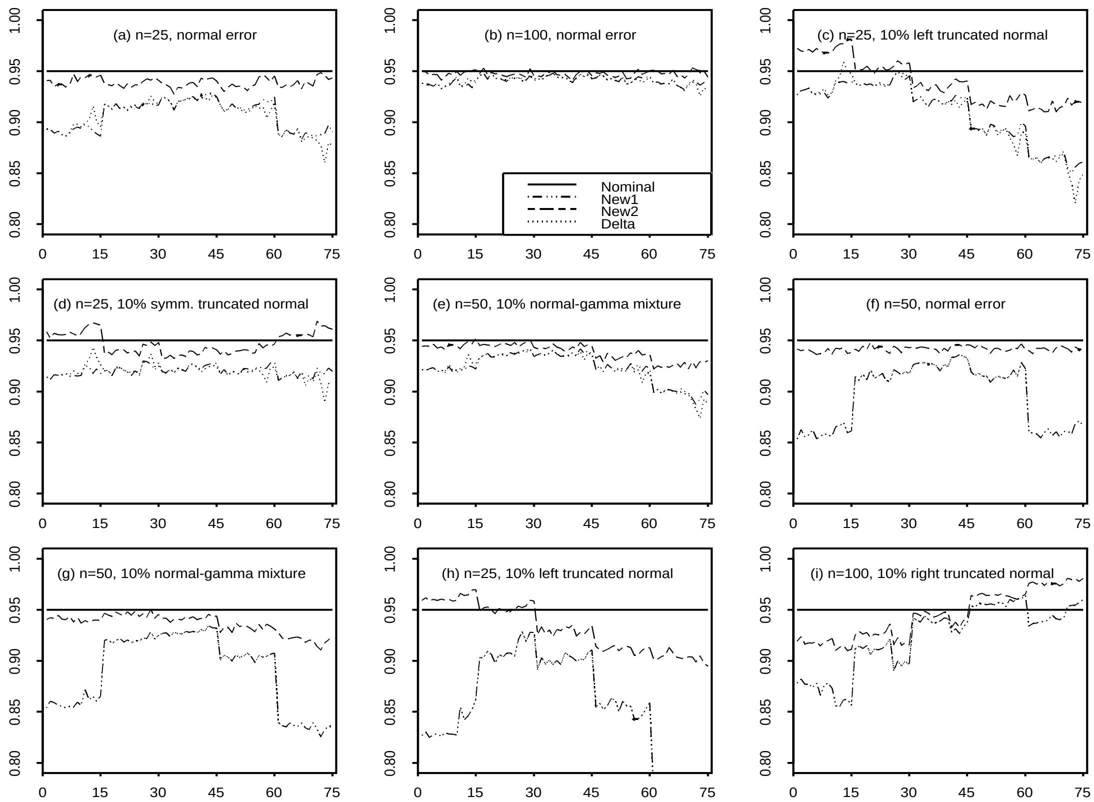

Figure 2: Coverage Probabilities of 95% CIs for Percentile Function ((a)-(e)), and Survivor Function ((f)-(i)), under Box-Cox Power Transformation

CI for survivor function. Simulation results presented in Figure 2 ((f)-(i)) and Figure 3(((e)-(i))) clearly demonstrate the point discussed in Section 3: the nonlinearity of the function g can have a large effect on the finite sample performance of the confidence intervals when applying the delta or equivalent methods. The results show that, when errors are normal, the finite sample performance of the CI (15) is much better than the other two CIs with all the coverage probabilities very close to 0.95, even when n is as small as 25. In contrast, the coverage probabilities of the other two CIs can be as low as 0.7238, and as n gets large they converge very slowly to 0.95, especially when p is small or large. It is well known that an accurate estimation of the tail probabilities is very important in the fields of reliability, medical research, actuarial science, insurance, etc. Thus, the CI (15) given in Section 3.5 should be recommended for the applications in these fields.

In the cases of truncated normal and normal-gamma mixture errors, similar patterns as for the percentile function are observed for the survivor function, except that the contrast between our Method 2 and the other two methods are much sharper for the case of the survivor function. Method 2 leads to CIs for the survivor function with reasonable finite sample performance in almost all situations studied, except when extreme tail probabilities are involved.

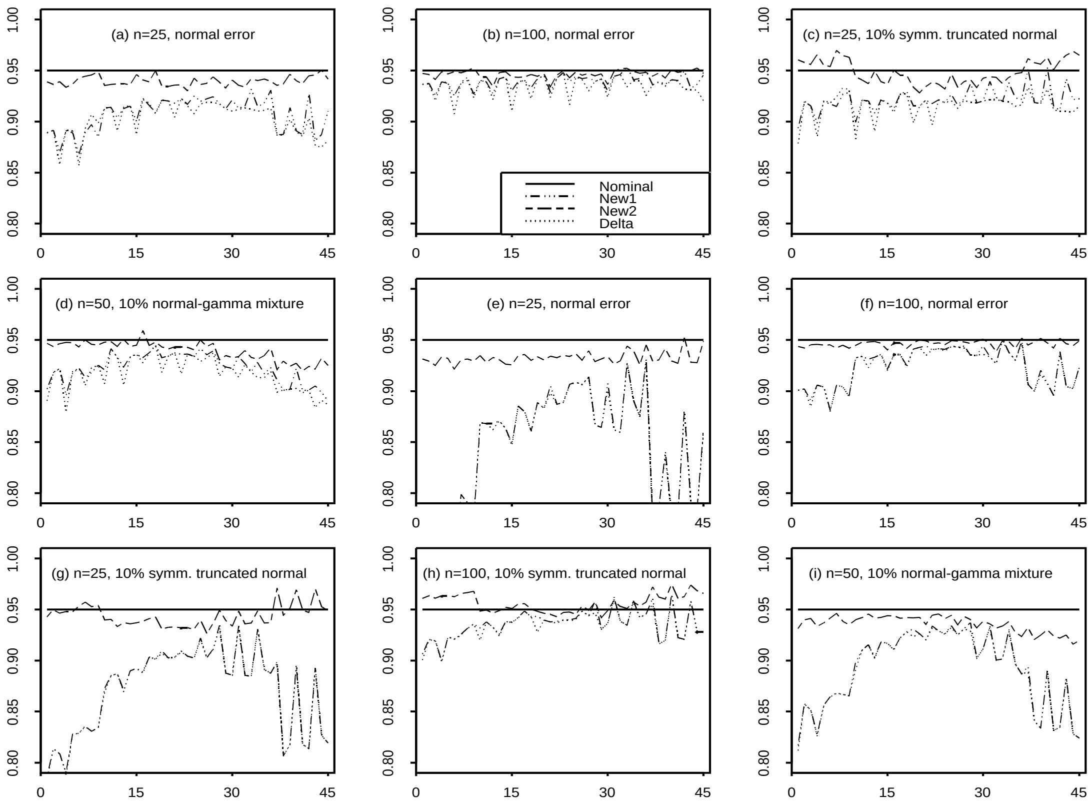

Figure 3: Coverage Probabilities of 95% CIs for Percentile Function ((a)-(d)), and Survivor Function ((e)-(i)), under Dual Power Transformation

In general, λ affects the coverage probability, but in a much smaller magnitude compared to the effects of p. The value of σ does not have much effect on the coverage probability. Finally, we have also recorded the average lengths of the CIs in the above simulations. As expected, the CIs based on our Method 2 are slightly longer than the CIs based on the other two methods.

5. AN EXAMPLE

The salary survey data of Chatterjee and Hadi (2006, p122), reproduced here in Table 1 for convenience, is used to illustrate our methods, in particular the CIs for percentile and survivor functions. The response variable is `Salary` and the predictors are years of experience (`Exp`), education (`Edu`) (1=high school diploma, 2=bachelor degree, 3=advanced degree), and management responsibility (`Man`) taking value 1 if a person bears management responsibility and 0 otherwise. The MLEs of the transformation parameter from the two transformations are, respectively, 0.183606 (Box-Cox) and 0.190988. The MLEs of the intercept, the coefficient of `Exp`, and the coefficients of three dummy variables (high school diploma, bachelor degree and `Man`) are (24.8645, 0.1913, -0.9647, 0.0367, 2.3575) from the Box-Cox power transformation, and (15.1719, 0.1053, -0.5308, 0.0202, 1.2974) from the dual power transformation. The MLEs of the error standard deviation are 0.3052 and 0.1679, respectively, from the Box-Cox power transformation and the dual power transformation.

Table 1: Salary Survey Data

Row	Salary	Exp	Edu	Man	Row	Salary	Exp	Edu	Man
1	13876	1	1	1	24	22884	6	2	1
2	11608	1	3	0	25	16978	7	1	1
3	18701	1	3	1	26	14803	8	2	0
4	11283	1	2	0	27	17404	8	1	1
5	11767	1	3	0	28	22184	8	3	1
6	20872	2	2	1	29	13548	8	1	0
7	11772	2	2	0	30	14467	10	1	0
8	10535	2	1	0	31	15942	10	2	0
9	12195	2	3	0	32	23174	10	3	1
10	12313	3	2	0	33	23780	10	2	1
11	14975	3	1	1	34	25410	11	2	1
12	21371	3	2	1	35	14861	11	1	0
13	19800	3	3	1	36	16882	12	2	0
14	11417	4	1	0	37	24170	12	3	1
15	20263	4	3	1	38	15990	13	1	0
16	13231	4	3	0	39	26330	13	2	1
17	12884	4	2	0	40	17949	14	2	0
18	13245	5	2	0	41	25685	15	3	1
19	13677	5	3	0	42	27837	16	2	1
20	15965	5	1	1	43	18838	16	2	0
21	12336	6	1	0	44	17483	16	1	0
22	21352	6	3	1	45	19207	17	2	0
23	13839	6	2	0	46	19346	20	1	0

Table 2 summarizes the CIs for percentile functions with $p = 0.05, 0.25, 0.5, 0.75$ and 0.95, and the CIs for survivor functions at y_0 chosen such that the values for the survivor function are estimated to be 0.95, 0.75, 0.5, 0.25, and 0.05, respectively. We choose $x_0' = (1, 10, 0, 0, 1)$, i.e. `Exp`=10, `Edu` = 3, and `Man`=1. The results show that the CIs based on Method 1 and the delta method are very similar. The CIs based on Method 2 are longest. These results are consistent with the Monte Carlo results – CIs based on Method 1 and the delta method often undercover the true quantity and as a result they are shorter. The results also show that the two transformations produce very similar sets of confidence intervals. Furthermore, a drawback of Method 1 and the delta method is clearly reflected in the CIs for the survivor function: the CIs based on these two methods can have an upper bound larger than 1 when y_0 is small, and a negative lower bound when y_0 is large. These problems do not occur with the corresponding CIs based on Method 2.

Table 2: CIs for percentile and survivor functions based on salary survey data

p or y_0	New Method 1		New Method 2		Delta Method	
Percentile Function, Box-Cox Power Transformation						
0.05	20981	22540	20705	22417	20970	22529
0.25	22081	23535	21929	23516	22072	23526
0.50	22819	24304	22755	24372	22810	24295
0.75	23524	25149	23548	25317	23513	25138
0.95	24484	26499	24634	26834	24468	26482
Percentile Function, Dual Power Transformation						
0.05	20981	22541	20705	22417	20969	22529
0.25	22081	23536	21929	23517	22072	23526
0.50	22819	24305	22755	24373	22810	24296
0.75	23525	25151	23549	25318	23514	25139
0.95	24486	26500	24636	26836	24469	26484
Survivor Function, Box-Cox Power Transformation						
21749	0.8742	1.0258	0.8186	0.9913	0.8742	1.0258
22799	0.5402	0.9595	0.5058	0.9088	0.5402	0.9595
23552	0.2383	0.7617	0.2559	0.7441	0.2383	0.7617
24325	0.0280	0.4723	0.0849	0.5099	0.0280	0.4723
25475	-0.0361	0.1360	0.0066	0.2088	-0.0361	0.1360
Survivor Function, Dual Power Transformation						
21749	0.8742	1.0258	0.8186	0.9913	0.8742	1.0258
22799	0.5402	0.9595	0.5058	0.9088	0.5402	0.9595
23553	0.2383	0.7617	0.2559	0.7441	0.2383	0.7617
24326	0.0280	0.4723	0.0849	0.5099	0.0280	0.4723
25477	-0.0360	0.1360	0.0066	0.2087	-0.0360	0.1360

6. CONCLUSIONS

We have provided general theories for conducting inferences, in a simple way, for the Box-Cox-type transformation model. More importantly, in the situations where the distribution of the λ-known statistic is completely known, we have introduced alternative procedures that lead to inference methods with better finite sample performance. Another important point as illustrated by the result (15) is that when the parametric function of interest is of limited range, a proper transformation of it may lead to a much better inference. Robustness of the methods are investigated by Monte Carlo simulation and the results show that our methods are quite robust against mild departures from normality of error distributions. No doubt, **simplicity**, **good finite sample performance**, and **robustness** are three important criteria for the applicability of a statistical inferential procedure. Our general results shed light in this direction and they may be extendable to the more complicated Box-Cox type of models.

APPENDIX

The detailed proofs of the theorems require the following set up and calculations. Let $\theta' = (\lambda, \beta', \sigma^2)$. Denote the partial derivatives of $h(Y_i, \lambda)$ with respect to Y_i and λ by adding relevant subscripts on h, i.e. $h_{y\lambda}(Y_i, \lambda) = \partial^2 h(Y_i, \lambda)/\partial Y_i \partial\lambda$. The score function $U(\theta) = \partial\ell(\theta)/\partial\theta$, partitioned according to λ, β, and σ^2, has elements,

$$
\begin{aligned}
U_1(\theta) &= -\frac{1}{\sigma^2}[h(\mathbf{Y},\lambda)-\mathbf{X}\beta]'h_\lambda(\mathbf{Y},\lambda)+\sum_{i=1}^{n}\frac{h_{y\lambda}(Y_i,\lambda)}{h_y(Y_i,\lambda)},\\
U_2(\theta) &= \frac{1}{\sigma^2}\mathbf{X}'[h(\mathbf{Y},\lambda)-\mathbf{X}\beta],\\
U_3(\theta) &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}[h(\mathbf{Y},\lambda)-\mathbf{X}\beta]'[h(\mathbf{Y},\lambda)-\mathbf{X}\beta].
\end{aligned}
$$

The Hessian matrix $H = \partial\ell^2(\theta)/\partial\theta\partial\theta'$ has elements according to λ, β and σ^2,

$$
\begin{aligned}
H_{11} &= -\frac{1}{\sigma^2}\{[h(\mathbf{Y},\lambda)-\mathbf{X}\beta]'h_{\lambda\lambda}(\mathbf{Y},\lambda)+h_\lambda(\mathbf{Y},\lambda)'h_\lambda(\mathbf{Y},\lambda)\}\\
&\quad +\sum_{i=1}^{n}\frac{h_{y\lambda\lambda}(Y_i,\lambda)h_y(Y_i,\lambda)-h_{y\lambda}^2(Y_i,\lambda)}{h_y^2(Y_i,\lambda)},\\
H_{22} &= -\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X},\\
H_{33} &= \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}[h(\mathbf{Y},\lambda)-\mathbf{X}\beta]'[h(\mathbf{Y},\lambda)-\mathbf{X}\beta],\\
H_{12} &= \frac{1}{\sigma^2}h_\lambda(\mathbf{Y},\lambda)'\mathbf{X},\\
H_{13} &= \frac{1}{\sigma^4}[h(\mathbf{Y},\lambda)-\mathbf{X}\beta]'h_\lambda(\mathbf{Y},\lambda),\\
H_{23} &= -\frac{1}{\sigma^4}\mathbf{X}'[h(\mathbf{Y},\lambda)-\mathbf{X}\beta].
\end{aligned}
$$

The expected information matrix ($-\text{E}[H(\theta)]$), denoted by $I(\theta)$ and partitioned as $I_{ij}, i,j = 1,2,3$, does not have explicit expressions for the elements involving subscript 1. For the other elements, we have $I_{22} = \sigma^{-2}\mathbf{X}'\mathbf{X}$, $I_{33} = \frac{n}{2\sigma^4}$, and $I_{23} = 0$. These expressions do not depend on which transformation function h the model employs. All the partial derivatives used in the theorems involving the Box-Cox power transformation, the dual power transformation, and the other transformations have analytical expressions and can all be derived easily.

Proof of Theorem 1. First, under regularity conditions for the ML estimation, we have

$$
\sqrt{n}\begin{pmatrix}\hat\beta(\lambda)-\beta\\ \hat\sigma^2(\lambda)-\sigma^2\end{pmatrix}\xrightarrow{D} N\left(0,\begin{pmatrix} n\sigma^2(\mathbf{X}'\mathbf{X})^{-1}, & 0\\ 0, & 2\sigma^4\end{pmatrix}\right).
$$

This leads immediately to $\sqrt{n}[\hat\psi(\lambda)-\psi]\xrightarrow{D} N(0,v^2)$. Next, the condition ii) and the first-order Taylor series expansion give

$$
\begin{aligned}
\sqrt{n}[\hat\psi(\hat\lambda)-\psi] &= \sqrt{n}[\hat\psi(\lambda)-\psi]+\hat\psi_\lambda(\lambda)\sqrt{n}(\hat\lambda-\lambda)+o_p(1)\\
&= \sqrt{n}[\hat\psi(\lambda)-\psi]+\kappa\sqrt{n}(\hat\lambda-\lambda)+o_p(1).
\end{aligned}
$$

Thus, the second result of the theorem follows by showing that the asymptotic covariance of $\sqrt{n}(\hat\psi(\lambda)-\psi)$ and $\sqrt{n}(\hat\lambda-\lambda)$ is zero, which can be done by i) expressing $\hat\psi$ linearly in $\hat\beta(\lambda)$ and $\hat\sigma(\lambda)$; ii) expressing $\hat\beta(\lambda)$ and $\hat\sigma(\lambda)$ asymptotically in terms of U_2, U_3 and $I_{ij}, i,j = 2,3$; and iii) expressing $\hat\lambda$ in terms of all elements of the score vector and the information matrix. Details of these calculations are tedious but available from the authors upon request.

Proof of Theorem 2. The first part is obvious. For the second part, condition iv) and a Taylor series expansion leads to

$$\begin{aligned}\sqrt{n}[c'\hat{\xi}(\hat{\lambda}) - f(\hat{\lambda},\psi)] &= \sqrt{n}[c'\hat{\xi}(\lambda) - f(\lambda,\psi)] + [c'\hat{\xi}_\lambda(\lambda) - f_\lambda(\lambda,\psi)]\sqrt{n}(\hat{\lambda}-\lambda) + o_p(1) \\ &= \sqrt{n}[c'\hat{\xi}(\lambda) - f(\lambda,\psi)] + \kappa\sqrt{n}(\hat{\lambda}-\lambda) + o_p(1).\end{aligned}$$

The second result of the theorem follows from similar calculations as in the proof of Theorem 1, considering the fact that $c'\hat{\xi}(\lambda)$ is a special case of $\hat{\psi}(\lambda)$.

The Delta Method. The delta method is simply a consequence of the following well-known result of ML estimation: if the MLE $\hat{\theta}$ is asymptotically normal with mean θ and variance-covariance matrix $I^{-1}(\theta)$, then a smooth function of $\hat{\theta}$, $g(\hat{\theta})$ say, is the MLE of $g(\theta)$ and is asymptotically normal with mean $g(\theta)$ and variance $g'_\theta(\theta)I^{-1}(\theta)g_\theta(\theta)$. The variance can be consistently estimated by $g'_\theta(\hat{\theta})J^{-1}(\hat{\theta})g_\theta(\hat{\theta})$, where $J = -H$ is the observed information matrix. Based on this result, one quickly obtains the statistic for making inference about ψ

$$T(\hat{\lambda},\psi) = \frac{\hat{\psi}(\hat{\lambda}) - \psi}{\sqrt{g'_\theta(\hat{\theta})J^{-1}(\hat{\theta})g_\theta(\hat{\theta})}}$$

with an asymptotic $100(1-\alpha)\%$ CI for ψ as

$$\hat{\psi}(\hat{\lambda}) \pm Z_{\alpha/2}\sqrt{g'_\theta(\hat{\theta})J^{-1}(\hat{\theta})g_\theta(\hat{\theta})}.$$

ACKNOWLEDGEMENTS

The authors are grateful to the Editor, Douglas Wiens, an Associate Editor and a referee for their constructive comments which have led to significant improvements in the paper. Z. L. Yang is grateful to the Department of Statistics, The Chinese University of Hong Kong and the Department of Mathematics and Statistics, University of Guelph for their hospitality during his visits. He gratefully acknowledges the research support from the Singapore Management University, and the excellent research assistance from Yihui Li. A. F. Desmond acknowledges with gratitude support via a Discovery Grant from the National Sciences and Engineering Research Council of Canada.

REFERENCES

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

G. E. P. Box & D. R. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society Series B*, 26, 211–252.

G. E. P. Box & D. R. Cox (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77, 209–210.

J. B. Burbidge, L. Magee & A. L. Robb (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83, 123–127.

G. Chen, R. A. Lockhart & M. A. Stephens (2002). Box-Cox transformations in linear models: large sample theory and tests of normality (with discussion). *The Canadian Journal of Statistics*, 30, 177–234.

R. J. Carroll (1982). Tests for regression parameters in power transformation models. *Scandinavian Journal of Statistics*, 9, 217–222.

R. J. Carroll & D. Ruppert (1984). Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, 79, 321–328

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

S. Chatterjee & A. S. Hadi (2006). *Regression Analysis by Example.* Hoboken, New Jersey: John Wiley & Sons. Inc.

A. M. Foster, L. Tian & L. J. Wei (2001). Estimation for the Box-Cox transformation model without assuming parametric error distribution. *Journal of the American Statistical Association*, 96, 1097–1101.

G. J. Hahn & W. Q. Meeker (1991). *Statistical Intervals, A Guide for Practitioners.* New York: John Wiley & Sons.

F. Hernandez & R. A. Johnson (1980). The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, 75, 855–861.

D. V. Hinkley & G. Runger (1984). The analysis of transformed data. *Journal of the American Statistical Association*, 79, 302–320.

J. A. John & N. A. Draper (1980). An alternative family of transformations. *Applied Statistics*, 29, 190–197.

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

B. F. Manly (1976). Exponential data transformation. *The Statistician*, 25, 37–42.

J. G. MacKinnon & L. Magee (1990). Transforming the dependent variable in regression models. *International Economic Review*, 31, 315–339.

W. M. Patefield (1977). On the maximized likelihood function. *Sankhya Series B*, 39, 92–96.

H. White (1994). *Estimation, Inference and Specification Tests.* Cambridge: Cambridge University Press.

Z. L. Yang (1999). Estimating a transformation and its effect on Box-Cox T-ratio. *Test*, 8, 167–190.

Z. L. Yang (2002a). Comment on " Box-Cox transformations in linear models: large sample theory and tests of normality ". *The Canadian Journal of Statistics*, 30, 222–225.

Z. L. Yang (2002b). Median estimation through a regression transformation. *The Canadian Journal of Statistics*, 30, 235–242.

Z. L. Yang (2006). A modified family of power transformations. *Economics Letters*, 92, 14–19.

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