The Canadian Journal of Statistics
Vol. 30, No. 2, 2002, Pages 1–59
La revue canadienne de statistique

Box–Cox transformations in linear models: large sample theory and tests of normality

Gemai CHEN, Richard A. LOCKHART and Michael A. STEPHENS

Key words and phrases: Empirical distribution function; goodness of fit; linear regression; maximum likelihood estimation; nonlinear regression; transformations to normality.

MSC 2000: Primary 62J05; secondary 62E20, 62G30.

Abstract: The authors provide a rigorous large sample theory for linear models whose endogenous variable has been subjected to the Box–Cox transformation. The theory provides a continuous asymptotic approximation to the distribution of natural estimates by focussing on the ratio of slope to standard deviation of the error term, since this parameter has a relatively stable and consistent estimate. The authors show the importance for inference of normality of the errors and give tests for normality based on the residuals. For nonnormal errors, the authors give adjustments to the log-likelihood and to asymptotic standard errors.

Les transformations de Box–Cox dans les modèles linéaires : théorie asymptotique et tests de normalité

Résumé : Les auteurs présentent une analyse asymptotique rigoureuse pour les modèles linéaires dont la variable endogène a été transformée par la méthode de Box–Cox. Partant du fait que le rapport entre la pente et l'écart-type du terme d'erreur admet des estimations convergentes et relativement stables, les auteurs sont conduits à une approximation asymptotique continue de la loi des estimateurs naturels du modèle. Ils montrent aussi l'importance du postulat de normalité des erreurs aux fins d'inférence et s'appuient sur les résidus pour vérifier ce postulat. Ils expliquent enfin comment ajuster la log-vraisemblance et les erreurs-type asymptotiques lorsque les termes d'erreur du modèle ne sont pas gaussiens.

1. INTRODUCTION

Table 1 shows 107 values of distance driven, $y_1, \ldots, y_{107}$ (in kilometers), and corresponding amounts of gasoline consumed, $x_1, \ldots, x_{107}$ (in liters). A natural, but naive, model for the relation between these variables is that $y_i = \beta x_i + \varepsilon_i$ for errors ε_i that are independent and identically distributed with mean 0 and variance σ^2.

This regression through the origin model is a submodel of the Box–Cox transformation model defined more generally as follows. Let Y be a positive random variable denoting a response to some covariates. For each real number λ, the Box–Cox transformation is

$$Y(\lambda) = \begin{cases} (Y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \log Y & \text{if } \lambda = 0. \end{cases} \tag{1}$$

The Box–Cox procedure selects a value λ so that after transformation, the following linear model is approximately applicable,

$$Y_i(\lambda) = \mu + x_i\beta + \sigma\varepsilon_i, \quad 1 \leq i \leq n \tag{2}$$

where $x_i = (x_{i,1}, \ldots, x_{i,p})$ is a known row vector of constants, $\beta = (\beta_1, \ldots, \beta_p)'$ is a column vector of unknown regression parameters, σ is an unknown positive constant, and ε_i is a random error with mean 0 and variance 1. We assume that the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed.

Box & Cox (1964) suggested that there will often be a single value of λ which achieves, or nearly achieves, three ends:

(i) the linear structure of the model will be simplified (for instance by removing interactions);

(ii) heteroscedasticity will be removed; and

(iii) the errors will be more nearly normal.

TABLE 1: Values of distance driven in kilometers y with corresponding liters of gasoline used x. Read down.

y	x	y	x	y	x	y	x	y	x
390.4	31.41	418.8	31.00	441.4	31.21	418.1	33.99	419.9	30.60
413.6	33.15	365.5	25.86	468.3	35.30	420.7	34.17	433.1	30.05
415.2	33.33	440.1	30.66	410.0	29.87	440.0	33.71	472.0	33.67
429.2	33.18	429.2	31.46	444.0	32.96	459.5	34.99	489.3	33.81
422.1	31.53	442.1	30.91	431.9	33.67	448.8	35.16	433.8	32.42
415.8	33.25	527.2	35.98	445.8	32.24	437.6	35.31	496.7	33.49
438.7	32.48	464.8	32.24	441.9	33.14	465.9	36.63	468.7	32.97
382.9	31.02	449.0	30.80	400.3	32.22	354.0	24.67	505.2	35.53
414.4	32.63	461.6	33.63	452.3	32.87	486.6	34.64	451.2	32.56
443.3	32.42	404.2	30.08	469.4	34.01	399.2	32.33	485.6	32.97
443.7	33.99	531.2	31.30	448.3	33.84	411.2	30.07	468.4	34.34
445.2	32.51	410.1	30.86	435.7	34.41	198.4	14.62	464.5	32.79
447.7	33.80	439.5	27.70	534.5	35.34	438.7	34.34	471.4	34.67
431.2	32.69	420.6	28.70	424.7	33.30	443.1	35.29	454.6	32.52
450.4	34.27	518.2	34.27	418.0	33.92	462.0	33.83	414.1	31.02
470.1	35.76	458.4	31.27	446.8	34.24	476.9	34.36	436.1	31.76
271.8	21.20	471.0	31.36	427.5	33.99	459.6	34.30	486.3	31.98
514.7	31.32	479.1	32.13	465.3	34.66	428.1	30.77	462.1	31.35
440.7	32.81	504.9	34.65	425.1	34.78	461.5	30.12	454.7	32.82
476.1	31.43	442.1	32.01	420.7	37.10	435.4	32.19	421.7	31.85
440.5	30.43	438.1	32.74	420.0	32.75	449.4	31.41	470.8	32.89
490.4	34.24	474.4	34.27						

Then, assuming that the errors ε_i are standard normal, Box and Cox proposed an estimate $\hat{\lambda}$ of λ based on maximum likelihood for the model (2). This can be found from a plot of the profile log-likelihood, $\ell_{BC}(\lambda)$; this is the logarithm of the maximum of the likelihood for a fixed λ. The estimate of λ will be $\hat{\lambda}$, the value which maximizes $\ell_{BC}(\lambda)$. For the data set in Table 1, the top left panel of Figure 1 shows the plot of $\ell_{BC}(\lambda)$; the estimate of λ can be seen to be $\hat{\lambda} \approx 1.46$, whereas $\lambda = 1$ corresponds to the naive model mentioned at the beginning of the Introduction. However, the latter value is contained in the usual 95% confidence interval for λ, whose endpoints meet $\ell_{BC}(\lambda)$ at a horizontal line (displayed in the figure) drawn a distance of $\chi^2_{1,0.95}/2$ below the maximum of $\ell_{BC}(\lambda)$.

When λ has been estimated by $\hat{\lambda}$, Box & Cox (1964) suggest that this value could be treated as being known and standard normal linear model theory applied to the set of $Y_i(\hat{\lambda})$ values to derive inferences for the regression parameters β and σ. Following such a suggestion for our data and using $\hat{\lambda} = 1.46$ would give $\hat{\beta} = 167.3 \pm 33$ (2 standard errors). When λ is unknown, however, the variability in $\hat{\beta}$ is much larger. In Figure 1, we have plotted the maximum likelihood estimates of the slope $\hat{\beta}(\lambda)$ and residual standard deviation $\hat{\sigma}(\lambda)$ as functions of λ. The estimates are plotted on a log scale; they vary by a factor of 10^4 over the range of statistically (as opposed to scientifically) plausible values for λ! Interestingly, however, the last panel of Figure 1 shows that the estimate $\hat{\theta}$ of the parameter $\theta = \beta/\sigma$, obtained as the ratio of the estimates of β and σ, varies only modestly over the range of λ. The parameters θ and $\delta = \lambda\sigma/(1 + \lambda\mu)$ play important roles in the analysis below.

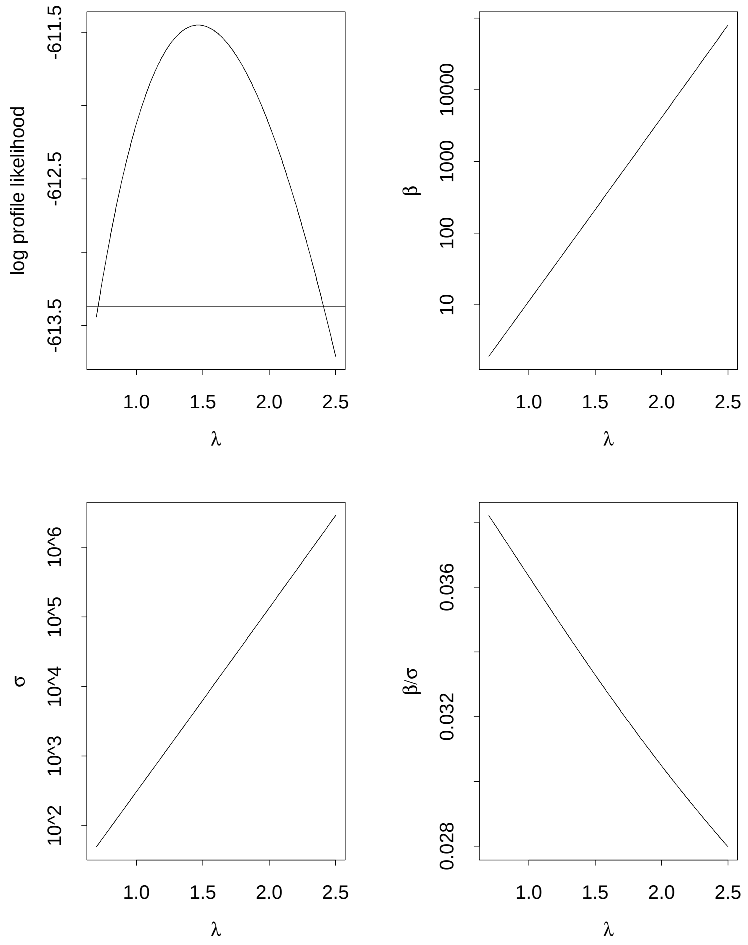

FIGURE 1: Top left: plot of the profile log-likelihood for λ; top right: the estimated slope $\hat{\beta}$ as a function of λ; bottom left: the estimated residual standard deviation $\hat{\sigma}$ as a function of λ; bottom right: the ratio $\hat{\theta} = \hat{\beta}/\hat{\sigma}$ as a function of λ. The symbolˆhas been omitted from the estimates $\hat{\beta}$, $\hat{\sigma}$ and $\hat{\theta}$.

1.1. Summary of the paper.

Our first goal in this paper is to make an asymptotic analysis to provide large sample approximations which predict the sort of plots seen in Figure 1, particularly the wide variability in $\hat{\beta}$ and in $\hat{\sigma}$, and the relative stability of $\hat{\theta}$ as λ is varied across its confidence interval.

It is well known that the Box–Cox model is explicitly approximate: some combinations of parameter values, covariate values and sample sizes do not describe data sets for which the model is reasonable. In particular, unless δ defined above is small, there will be a nonnegligible probability that the right-hand side of (2) takes on a value which is not a possible value of the left-hand side. In Section 2, we investigate some algebra related to the model, fixing n and taking a limit where δ does not remain constant but $\delta \to 0$, and the truncation effects can be ignored. Similar expansions for small values of the parameter δ have been used by Draper & Cox (1969) and by Taylor (1985a, 1985b and 1986).

In Section 3, we let $n \to \infty$ and $\delta \to 0$ simultaneously and examine asymptotic expansions of

$$\hat{\phi} \equiv \delta(\hat{\lambda} - \lambda)/\lambda$$

and of $\hat{\theta}$. We show that $\hat{\phi}$ has a limit as $\delta \to 0$ provided that θ is held fixed as the limit is taken. The usual Box–Cox estimates $\hat{\beta}$ and $\hat{\sigma}$ can be written in the forms $\hat{\beta} = A\tilde{\beta}$ and $\hat{\sigma} = A\tilde{\sigma}$, where $A = (1 + \lambda\mu)^{\hat{\phi}/\delta}$, and where $\tilde{\beta}$ and $\tilde{\sigma}$ each have reasonable asymptotic expansions. The presence of $1/\delta$ in the exponent in the coefficient A can produce very volatile, badly biased, estimates of β and σ. However, often $\hat{\theta} = \hat{\beta}/\hat{\sigma} = \tilde{\beta}/\tilde{\sigma}$ can be shown to be only slightly more variable than it would be if the true value of λ were known. Similar conclusions are in Doksum & Wong (1983) and Carroll (1982), at least in a testing context. They appear to be justified by the plots in Figure 1.

Tests and confidence intervals for λ and for $a'\theta$, where a is a column vector of constants, are discussed in Section 4. Theorems are given which justify, for normal errors, standard procedures for λ based on the log-likelihood; for example, the use of χ_1^2 to obtain the confidence band shown in Figure 1. Adjustments to the log-likelihood are given so that similar procedures may be used even with nonnormal errors.

The expansions for $\hat{\theta}$ lead to an asymptotic covariance matrix $\mathbf{V}$ of $\hat{\theta}$, and conditions are given to obtain asymptotic normality of $a'\hat{\theta}$. For the special case when the errors ε_i are normal, $\mathbf{V}$ simplifies to three terms; the first derives from the regular linear model theory, and the other two describe the effects of estimating λ and σ. More generally, asymptotic standard errors of components of $\hat{\theta}$ depend on the first six moments of the underlying error distribution.

In Sections 3 and 4, the assumption of normal errors simplifies many of the results; it is also important for the analysis of regression diagnostics which use the residuals. We give, in Section 5, goodness-of-fit tests of normality for the errors; these are based on the empirical distribution function of the residuals. In particular, the Anderson–Darling and Cramér–von Mises tests are developed in detail, and tables are given with which to make the test.

In Section 6, examples are given to show the use of our results in data analysis. Section 7 has a small Monte Carlo study suggesting that the asymptotic approximations are adequate for samples of reasonable size in situations likely to arise in practice.

Before turning to the expansions and tests, we summarize some of the work done by previous authors, and relate our work to theirs.

1.2. Background.

The suggestion of Box & Cox (1964) that $\hat{\lambda}$ be treated as the true value for the purpose of further estimation of parameters has led to sharp controversy over the question of variance inflation in the estimates of these parameters; for example, see Bickel & Doksum (1981), Box & Cox (1982) and Hinkley & Runger (1984).

Bickel & Doksum (1981) made asymptotic calculations showing that if the Box–Cox suggestion is followed, the usual estimate $\hat{\beta}$ of β is asymptotically more variable than standard linear model theory gives, and that this variance inflation effect can be very serious. However, the parameter β is measured in units which are those of Y to the power λ divided by the units of the covariates. In Figure 1, for instance, as λ varies from 1 to 2.5 the units of the slope $\hat{\beta}(\lambda)$ vary from km/L to km$^{2.5}$/L. In particular, $\hat{\beta}$ and the true β (corresponding to the true λ) will have different units. Box & Cox (1982) and Hinkley & Runger (1984) therefore argued that while Bickel and Doksum were mathematically correct to assert that $\hat{\beta} - \beta$ is more variable than standard linear model theory suggests, they were wrong to attach any meaning to this fact, since the subtraction of quantities with different units could not be scientifically meaningful.

Similar criticism will apply to the difference $\hat{\sigma} - \sigma$. Hinkley and Runger proposed that $\hat{\beta}$ be regarded as an estimate of a data dependent parameter $\mathrm{E}\,(\hat{\beta} \mid \hat{\lambda})$; the usual confidence intervals would be based on the variability of $\hat{\beta} - \mathrm{E}\,(\hat{\beta} \mid \hat{\lambda})$. See Hooper & Yang (1997) for an asymptotic analysis of this suggestion.

Our asymptotic analysis of the model (2) uses a limit which differs from that used by other authors. Since the difference is crucial to the motivation of this paper, we take some time to justify our framework. Asymptotic calculations are intended to provide approximations to similar calculations for fixed n; the approximations arise by regarding the model for the actual data collection as the nth in a sequence of models. Different sequences are obtained by making different assumptions on the parameters as $n \to \infty$; since the limits depend on the sequence, one must choose a sequence which appears to fit the problem at hand.

It is standard in statistical theory to assume that all the models in the sequence have the same parameter values. One can then imagine that data points are generated one after the other, the model for each new data point having the same parameter values as for all previous data points.

We shall depart from this assumption. We are analyzing a model for some particular value of n, and in the sequence of models, the parameter values λ, β and σ will be allowed to depend on n. We argue that a good sequence of models should have computable limiting distributions for estimates and should have finite sample distributions which converge quickly to their limits. The best sequence need not have constant parameter values.

In Section 3 and 4, we consider limits as $\delta \to 0$. Bickel & Doksum (1981) also take a limit in which $\delta \to 0$ as $n \to \infty$; they hold λ and β fixed but make $\sigma \to 0$ so that $\delta \to 0$. The implication is then that a larger sample has all the data coming from a more accurate model (2) than from a smaller sample. The limit computed by Bickel and Doksum behaves discontinuously at $\beta = 0$. In the Box–Cox model, the joint density of the data depends continuously on the parameter values when n is fixed. In any such situation, limit distributions which depend discontinuously on the parameters must provide bad approximations for any n for some parameter values. In the case of the Bickel and Doksum asymptotic calculations, these bad approximations occur for values of β near 0, values which can easily arise in practice.

In order to avoid the discontinuity, we focus on $\theta = \beta/\sigma$ and embed our model in a sequence of models for which θ is approximately fixed as $n \to \infty$. One achieves this either by fixing β and allowing $\mu \to \infty$ or $\lambda \to 0$, or by taking $\beta \to 0$ with $\sigma \to 0$. In either case, we get a limit which depends continuously on θ. We argue that such a limit provides a better description of the impact on inference of using an estimated value of λ since it removes the discontinuity at $\theta = 0$. Our results show that while the estimates $\hat{\beta}$ and $\hat{\sigma}$ can be subject to gross variance inflation, the ratio $\hat{\theta} \equiv \hat{\beta}/\hat{\sigma}$ is much less sensitive to estimation of λ.

In addition to being estimable in a stable way, θ has the important advantage over β that the units are constant; θ is simply measured in inverse units of the covariates. In the example the units are inverse liters. This makes the subtraction $\hat{\theta} - \theta$ physically meaningful. Consideration of θ permits us to avoid both data dependent parameters, such as $\mathrm{E}\,(\hat{\beta} \mid \hat{\lambda})$, and the subtraction of incomparables.

Consideration of the parameter θ has been suggested by a number of authors. For instance, Li & Duan (1989) worked in a more general transformation context and argue for consideration of $\beta_i/\beta_j = \theta_i/\theta_j$. Cox & Reid (1987) studied a reparametrization of the problem for which the estimate of λ is orthogonal to estimates of a number of other parameters at $\beta = 0$ and which leads to consideration of θ. Our results may be regarded as providing a rigorous asymptotic justification for the approach of Cox and Reid. Moreover, we identify several key quantities that can be estimated to assess the extent of the variance inflation problem when $\beta \neq 0$.

An important feature of Section 3 is that we do *not* show that the estimates of β and σ are consistent nor that the estimate of the transformation parameter λ is consistent. Only the parameter θ is, in general, estimated consistently. This conclusion contrasts sharply with earlier work on large sample theory. See Hernandez & Johnson (1980) and Carroll (1982), Carroll &

Ruppert (1984) as well as Bickel & Doksum (1981) for a large sample framework in which the estimate of λ is consistent.

2. THE LIMIT OF NEGLIGIBLE TRUNCATION

In this section, we study the limit when n is held fixed and δ is permitted to converge to 0. Our goal is to discover those statistics which are fairly stable, that is, whose behaviour is not dramatically dependent on the value of δ.

Fix n. Add a subscript 0 to distinguish true parameter values λ_0, β_0, σ_0 and θ_0 from arguments of functions such as λ in the log-likelihood. Let

$$Z_i = \mu_0 + \sigma_0(\eta_i + \varepsilon_i),$$

where η_i = $x_i'\theta_0$, $\theta_0 = \beta_0/\sigma_0$. Our data will be $Y_i = (1+\lambda_0 Z_i)^{1/\lambda_0}$; for $\lambda_0 = 0$ this becomes $Y_i = \exp(Z_i)$. Note that the Y_i, thus defined, satisfy the Box–Cox model (2). For $\lambda_i > 0$, if $Z_i < -1/\lambda_0$ put $Y_i = 0$; similarly for $\lambda_i < 0$, if $Z_i > -1/\lambda_0$ put $Y_i = 0$. We will be taking a limit in which the probability tends to 0 that any Y_i is 0.

It will be convenient to use vector notation. For the complete data set with n values of Y_i, we write $\mathbf{Y}(\lambda) = (Y_1(\lambda), \ldots, Y_n(\lambda))'$, where $'$ denotes transposition. Without loss of generality, we can suppose x_i has been centered; that is, replace the original x_i by $x_i - (x_1 + \cdots + x_n)/n$. Then the $n \times (p+1)$ design matrix, $\mathbf{X}$, has the form $\mathbf{X} = (\mathbf{1}|\mathbf{X}_1)$, where $\mathbf{1}$ is a column vector of 1's, and the $n \times p$ matrix $\mathbf{X}_1$ has ith row the centered x_i. Then $\mathbf{1}'\mathbf{X}_1 = 0$. Let $\mathbf{M} = \mathbf{I} - \mathbf{H}$, where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the usual hat matrix. Finally let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. The conventional form of the linear model is $\mathbf{Y}(\lambda) = \mu\mathbf{1} + \mathbf{X}_1\beta + \sigma\varepsilon$.

Treating the likelihood for (2) as if the errors were normal and then maximizing over μ, β and σ leads to the log-profile-likelihood for λ:

$$\ell_{BC}(\lambda) = -\frac{n}{2}\log\bigl\{\mathbf{Y}(\lambda)'(\mathbf{I}-\mathbf{H})\mathbf{Y}(\lambda)/n\bigr\} - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^{n}\log(Y_i). \tag{3}$$

With true parameters, δ becomes

$$\delta = \lambda_0\sigma_0/(1+\mu_0\lambda_0).$$

Define

$$\phi = \delta(\lambda - \lambda_0)/\lambda_0.$$

For $\lambda_0 = 0$, put $\phi = \sigma_0\lambda$. We have

$$Y_i^\lambda = (1+\lambda_0\mu_0)^{1+\phi/\delta}\bigl\{1+\delta(\eta_i+\varepsilon_i)\bigr\}^{1+\phi/\delta}. \tag{4}$$

Then consider the log-likelihood ratio $\ell(\phi) \equiv \ell_{BC}(\lambda) - \ell_{BC}(\lambda_0)$ for an assumed value λ. Use $\mathbf{M1} = 0$ to check that

$$\ell(\phi) = -\frac{n}{2}\log\left\{\frac{U'(\phi,\delta)\mathbf{M}U(\phi,\delta)}{\varepsilon'\mathbf{M}\varepsilon}\right\} + \frac{\phi}{\delta}\sum_{i=1}^{n}\log\bigl\{1+\delta(\eta_i+\varepsilon_i)\bigr\}, \tag{5}$$

where $U(\phi,\delta)$ is the vector with ith coordinate

$$U_i(\phi,\delta) = \frac{\bigl\{1+\delta(\eta_i+\varepsilon_i)\bigr\}^{1+\phi/\delta} - 1}{\phi+\delta}.$$

We now record various estimates in terms of U. The formulas show which quantities can be evaluated stably when δ is small and which cannot. Let $\hat\phi$ maximize (5). Let

$$\tilde\beta(\phi) = \sigma_0(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'U(\phi,\delta) \quad \text{and} \quad \tilde\sigma^2(\phi) = \frac{\sigma_0^2 U(\phi,\delta)'\mathbf{M}U(\phi,\delta)}{n-p-1}.$$

The estimated regression vector is $\hat{\beta} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}(\hat{\lambda})$, and using (4) we have

$$\hat{\beta} = \sigma_0(1+\lambda_0\mu_0)^{\hat{\phi}/\delta}(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'U(\hat{\phi},\delta) = (1+\lambda_0\mu_0)^{\hat{\phi}/\delta}\tilde{\beta}(\hat{\phi}). \tag{6}$$

The estimated residual standard deviation is given by $\hat{\sigma} = (1+\lambda_0\mu_0)^{\hat{\phi}/\delta}\tilde{\sigma}$. Thus the usual estimate of the standardized regression parameters is $\hat{\theta} = \hat{\beta}/\hat{\sigma} = \tilde{\beta}/\tilde{\sigma}$, where $\tilde{\beta} = \tilde{\beta}(\hat{\phi})$ and $\tilde{\sigma} = \tilde{\sigma}(\hat{\phi})$. The vector of fitted residuals is

$$\sigma_0(1+\lambda_0\mu_0)^{1+\hat{\phi}/\delta}\mathbf{M}U(\hat{\phi},\delta).$$

The standardized residuals are $\hat{\varepsilon} = \mathbf{M}U(\hat{\phi},\delta)/(\tilde{\sigma}/\sigma_0)$.

We are now ready to take the limit as $\delta \to 0$ with n fixed. Note that the limit $\delta \to 0$ can arise in one of three ways: $\lambda_0 \to 0$, $\mu_0 \to \infty$, or $\sigma_0 \to 0$. Note that the limit of $U(\phi,\delta)$ as $\delta \to 0$ is

$$U_i(\phi,0) = \frac{\exp\{\phi(\eta_i+\varepsilon_i)\}-1}{\phi}.$$

Then $\ell(\phi)$ has limit $\ell_0(\phi)$, where

$$\ell_0 = -\frac{n}{2}\log\left\{\frac{U(\phi,0)'\mathbf{M}U(\phi,0)}{\varepsilon'\mathbf{M}\varepsilon}\right\} + \phi\sum_{i=1}^{n}\varepsilon_i.$$

Thus as $\delta \to 0$, the vector of standardized residuals $\hat{\varepsilon}$ converges to $\mathbf{M}U(\hat{\phi}_0,0)/\hat{\sigma}(\hat{\phi}_0)$, where $\hat{\phi}_0$ minimizes ℓ_0 and $n\hat{\sigma}^2(\hat{\phi}_0) = U(\hat{\phi}_0,0)'\mathbf{M}U(\hat{\phi}_0,0)$. Moreover, the empirical process Y_n derived from the standardized residuals converges (weakly in $D[0,1]$) to that derived from the entries in $\mathbf{M}U(\hat{\phi}_0,0)/\hat{\sigma}(\hat{\phi}_0)$.

3. ASYMPTOTIC EXPANSIONS OF ESTIMATES

3.1. Expansion of $\hat{\phi}$.

In the previous section, n was fixed. Now we let $n \to \infty$ and in addition we let $\delta \to 0$ in order to obtain asymptotic results. We concentrate on the necessary conditions and the conclusions, and leave the proofs of the theorems to Appendix A.

Our asymptotic calculations follow the standard framework: representation of ϕ as $-\ell'(0)/\ell''(0)$, followed by expansion of $\ell'(0)$ and $\ell''(0)$. In the case of normal errors, we demonstrate that the variance of $\ell'(0)$ is approximately $-\ell''(0)$. We are then able to provide asymptotic expansions of the estimate $\hat{\theta}$.

In Chen, Lockhart & Stephens (2001), we give much weaker conditions than we do here. Those conditions highlight the relations between δ, the leverages H_{ii} and the moments of the set of effects η_i. They also permit the parameter value θ to grow with n, providing some overlap with the Bickel and Doksum framework. In this paper, we simplify the presentation by strengthening substantially the conditions.

Though later conditions will often be more stringent, we will need

$$\delta \max_{1\le i\le n}|\eta_i| \to 0. \tag{7}$$

Note that (7) imposes a condition on $\mathbf{X}$ and θ together; the condition is not generally satisfied in the Bickel and Doksum asymptotics.

Our work permits the dimension p of the parameter θ to grow with the sample size. We will need the basic assumption

$$p^3/n \to 0. \tag{8}$$

We impose a further constraint on the rate of growth of p; the condition permits p to grow more quickly when δ goes to 0 more quickly with n:

$$\delta^4 p \to 0. \tag{9}$$

We require moment conditions on the η_i and the ε_i:

$$\limsup_{n\to\infty} \sum_{i=1}^{n} \eta_i^8/n < \infty \tag{10}$$

and, letting μ_j denote the jth (central) moment of the ε_i,

$$\mu_8 < \infty. \tag{11}$$

The following conditions permit $\delta \to 0$ more slowly for errors which are more nearly normal. We assume either (A1) or (A2):

(A1) the errors have a standard normal distribution and $\delta^2 \log(n) \to 0$;

(A2) there is an integer $q \geq 3$ such that the following three conditions hold:

(A2.1) the errors have the same first q moments as the standard normal distribution,

$$\mu_j \equiv \mathrm{E}\,(\varepsilon^j) = \begin{cases} 0 & \text{if } j \text{ is odd and } j \leq q, \\ (j-1)(j-3)\cdots 3 & \text{if } j \text{ is even and } j \leq q. \end{cases}$$

(A2.2) $\mathrm{E}\,|\varepsilon|^{2q} < \infty$.

(A2.3) $\delta^{2q-4} n \to 0$.

Note that condition (A2.1) implies that the errors have zero skewness, i.e., $\mathrm{E}\,(\varepsilon^3) = 0$. Without this condition, the profile score, $\ell'(0)$, is biased even when $\delta = 0$ (as happens when the true value of λ is 0) and none of the parameter estimates will be consistent. Note also that (A2.2) is stronger than (11) for $q \geq 5$. Finally, note that either (A1) or (A2) implies

$$\delta \max_{1\leq i\leq n} |\varepsilon_i| = o_P(1). \tag{12}$$

The above conditions permit the usual conclusions about the existence and asymptotic normality of a consistent root of the likelihood equations. We shall use the notation

$$g = \frac{3}{2}n + 2\sum_{i=1}^{n} \eta_i^2 + \frac{1}{4}(\eta^{\#2})'\mathbf{M}(\eta^{\#2}),$$

$$g^* = \left\{(\mu_6 - 6\mu_4 + 9)n + 4(\mu_4 - 1)\sum_{i=1}^{n} \eta_i^2 + (\eta^{\#2})'\mathbf{M}(\eta^{\#2})\right\}\Big/4$$

and

$$g^{**} = \left\{(7\mu_4/3 - 1)n + 8\sum_{i=1}^{n} \eta_i^2 + (\eta^{\#2})'\mathbf{M}(\eta^{\#2})\right\}\Big/4,$$

where $\eta_i = x_i\theta$ and $\eta^{\#2}$ is the column vector with ith entry η_i^2. Define also the vector $r = \mathbf{M}\eta^{\#2}$; entries are r_i. It will be convenient to set

$$\Omega_i = \varepsilon_i(3 - r_i)/2 - \eta_i(\varepsilon_i^2 - 1) - \varepsilon_i^3/2.$$

THEOREM 1. *Assume* (7), (8), (9) (10), (11) *and either* (A1) *or* (A2). *Then there is a sequence* L_n *increasing to infinity so slowly that on an event whose probability approaches* 1 *the supremum of the log-profile likelihood* ℓ *over the set* $|\phi| \le L_n/\sqrt{g}$ *is achieved at a point* $\hat\phi$ *which is the unique solution of* $\ell'(\phi) = 0$ *over the set in question. This point* $\hat\phi$ *has the asymptotic expansion*

$$\hat\phi = \sum \Omega_i/g^{**} + o_P\big(1/\sqrt{n}\big)$$

and

$$\frac{g^{**}\hat\phi}{\sqrt{g^*}} = \frac{g^{**}\delta(\hat\lambda - \lambda)}{\lambda\sqrt{g^*}} \Rightarrow \mathsf{N}(0,1).$$

For normal errors (or errors with the same first six moments as normal errors), we have $g = g^* = g^{**}$. *Then*

$$g^{1/2}\hat\phi = \frac{g^{1/2}\delta(\hat\lambda - \lambda)}{\lambda} \Rightarrow \mathsf{N}(0,1).$$

3.2. Expansions of $\hat\theta$.

We now turn to expansions of other estimates. We have not yet given conditions which would make the asymptotic standard errors of our estimates have limits. Our expansions are phrased in terms of the distribution of an arbitrary linear combination of the entries in $\hat\theta$; then when the dimension p of θ is fixed, the Cramér–Wold device will give multivariate limiting distributions. Recall that $\tilde\beta(\phi) = \sigma(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'U(\phi,\delta)$ and $\tilde\sigma^2(\phi) = \sigma^2 U(\phi,\delta)'\mathbf{M}U(\phi,\delta)/(n-p-1)$. Put $v = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\eta^{\#2}$. Then:

THEOREM 2. *Under the conditions of Theorem* 1,

$$\sup\left|\frac{\tilde\sigma^2(\phi)}{\sigma^2} - 1 - \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i^2 - 1)\right| = o_P\big(1/\sqrt{n}\big). \tag{13}$$

Let $a = a_n$ *be a sequence of vectors of dimension* p. *Then*

$$\sup\big\{|a'\tilde\beta(\phi)/\sigma - a'\theta - a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon - \phi a'v/2|\big\} = o_P\Big(\sqrt{a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a}\Big). \tag{14}$$

If the sequence $a_n'\theta$ *is bounded, then*

$$\begin{aligned}\sup\left\{\left|a'\hat\theta(\phi) - a'\theta - a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon - \frac{\phi a'v}{2} + \frac{a'\theta}{2n}\sum_{i=1}^{n}(\varepsilon_i^2 - 1)\right|\right\}\\ = o_P\big(1/\sqrt{n}\big) + o_P\Big(\sqrt{a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a}\Big).\end{aligned} \tag{15}$$

Finally,

$$\sup\big[\{\hat\theta(\phi) - \theta\}'\mathbf{X}_1'\mathbf{X}_1\{\hat\theta(\phi) - \theta\}\big] = \sup\big[\{\hat\eta(\phi) - \eta\}'\{\hat\eta(\phi) - \eta\}\big] = O_P(p). \tag{16}$$

In the expansions, the suprema are over $|\phi| \le L/\sqrt{n}$ *for any fixed* $L > 0$.

Equation (15) implies the following expansion of $\hat\theta$:

$$\hat\theta(\phi) \approx \theta + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon + \frac{\phi v}{2} - \frac{\theta}{2n}\sum_{i=1}^{n}(\varepsilon_i^2 - 1).$$

If we replace ϕ by $\hat\phi$ and apply Theorem 1, we get the expansion

$$\hat\theta \approx \theta + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon - \frac{\theta}{2n}\sum_{i=1}^{n}(\varepsilon_i^2 - 1) + \frac{v}{2g^{**}}\sum_{i=1}^{n}\Omega_i.$$

This gives an asymptotic covariance matrix $\mathbf{V}$ of $\hat{\theta}$ of the form

$$\mathbf{V}(\hat{\theta}) = (\mathbf{X}_1'\mathbf{X}_1)^{-1} + \frac{\theta\theta'(\mu_4 - 1)}{4n} + \frac{g^*}{4(g^{**})^2}vv' + \frac{\mu_5}{8g^{**}}(v\theta' + \theta v').$$

Covariance matrices are usually used to provide confidence intervals for quantities of the form $a'\theta$, provided the estimate $a'\hat{\theta}$ is approximately normally distributed. Our conditions are not strong enough to guarantee asymptotic normality of $a'\hat{\theta}$ for an arbitrary sequence a. Adding the standard condition $\max(H_{ii}) \to 0$ gives asymptotic normality of $a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon$, which then implies asymptotic normality of $a'\hat{\theta}$.

For the special case of normal errors, we have $\mu_5 = 0$ and $g^* = g^{**}$, so that

$$\mathbf{V}(\hat{\theta}) = (\mathbf{X}_1'\mathbf{X}_1)^{-1} + \frac{1}{2n}\theta\theta' + \frac{vv'}{4g}. \tag{17}$$

Moreover, in the normal case, the further condition $\max(H_{ii}) \to 0$ is not needed to guarantee the asymptotic normality of $a'\hat{\theta}$.

More generally, the first two terms in the expansion of $\hat{\theta}$ give the exact representation of $\hat{\beta}/\sigma$ in the usual linear model context which would arise if λ were known. The other two terms represent effects due to estimation of ϕ (i.e., λ), and of σ and are both negligible if $\theta = 0$.

In many cases, $\mathbf{V}$ will differ only slightly from the classical $(\mathbf{X}_1'\mathbf{X}_1)^{-1}$. Some examples are given in Section 6. Notice that, as in Cox & Reid (1987), there is no variance inflation if $\beta = 0$. Indeed there is little variance inflation if a regression of the squared effects η_i^2 on the original design matrix produces negligibly small regression coefficients v.

As an example, consider a simple linear regression with the covariate values x_i spread at equal intervals or in some way arranged so that $\sum x_i^3 = 0$. Then $v = 0$. Similarly, a two-sample problem may be regarded as a linear model by taking x_i as a dummy variable. If the two sample sizes are equal, the covariate values x_i will be 1 for observations in one of the samples and -1 for observations in the other sample. Then $\sum x_i^3 = 0$ and $v = 0$.

However, it is central to the conclusion that θ and not β is being estimated. Moreover, it is important to remember that estimation of σ will inflate the variance of $\hat{\theta}$ whenever the regression coefficients are not 0.

3.3. Variance estimation.

The asymptotic variances of $\hat{\lambda}$ and $a'\hat{\theta}$ can be estimated consistently by replacing all unknown parameters with corresponding estimates. Specifically, estimate μ_r, for $r = 4$, 5 or 6 using $\hat{\mu}_r = \sum \hat{\varepsilon}_i^r/n$. Estimate η using $\hat{\eta} = \mathbf{X}\hat{\theta}$ and get $\hat{g}$, $\hat{g}^*$ and $\hat{g}^{**}$ by replacing all occurrences of μ_r or η by the corresponding estimate. Similarly, the estimated variance of $a'\hat{\theta}$ is $a'\widehat{\mathbf{V}}a$, where $\widehat{\mathbf{V}}$ is $\mathbf{V}$ with μ_r replaced by $\hat{\mu}_r$, θ replaced by $\hat{\theta}$ and $v = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\eta^{\#2}$ replaced by $\hat{v} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\hat{\eta}^{\#2}$. In Appendix A, we prove the following theorem.

THEOREM 3. *Assume the conditions of Theorem 1. Assume that the support of the error distribution has cardinality greater than* 3. *Then*

$$\hat{g}/g = 1 + o_P(1), \quad \hat{g}^*/g^* = 1 + o_P(1), \quad \hat{g}^{**}/g^{**} = 1 + o_P(1).$$

Furthermore, $a'\widehat{\mathbf{V}}a/a'\mathbf{V}a = 1 + o_P(1)$ *and* $\hat{\mu}_r - \mu_r = o_P(1)$ *for* $r = 4, 5, 6$.

The second assumption, which rules out only a few pathological distributions for the errors, is required only for the last result.

4. HYPOTHESIS TESTS AND CONFIDENCE INTERVALS

The theorems above can be applied to give tests and confidence intervals for λ and $a'\theta$.

4.1. Parameter λ.

Consider the null hypothesis $\lambda = \lambda_0$. When this null hypothesis is true, we see that

$$2\ell(\hat{\phi}) = 2\{\ell_{BC}(\hat{\lambda}) - \ell_{BC}(\lambda_0)\}.$$

Theorem 1 shows that, for the case of normal errors, when $g^* = g^{**} = g$, likelihood-based confidence intervals for λ, using the χ_1^2 approximation to the null distribution of the test statistic $2\ell(\hat{\phi})$, are asymptotically justified. More generally this log-likelihood ratio must be replaced by

$$Z = 2\hat{g}^{**}\{\ell_{BC}(\hat{\lambda}) - \ell_{BC}(\lambda_0)\}/\hat{g}^*.$$

Critical points and p-values are then obtained by referring $2\ell(\hat{\phi})$ or Z to the χ_1^2 distribution.

The asymptotic standard error of $\hat{\lambda}$ is $\sigma_{\hat{\lambda}} \equiv |\lambda|\sqrt{g^*}/(|\delta|g^{**})$. Under our conditions, this quantity need not go to 0; thus $\hat{\lambda}$ need not be consistent. However, we can show that $(\hat{\lambda}/\lambda)/(\hat{\delta}/\delta) \to 1$. In view of Theorem 3, if we define $\hat{\sigma}_{\hat{\lambda}} = |\hat{\lambda}|\sqrt{\hat{g}^*}/(|\hat{\delta}|\hat{g}^{**})$, then we have $\hat{\sigma}_{\hat{\lambda}}/\sigma_{\hat{\lambda}} \to 1$ in probability. Under the assumption of normal errors, the simpler estimate $\hat{\sigma}^*_{\hat{\lambda}} = |\hat{\lambda}/(\hat{\delta}\hat{g}^{1/2})|$ also satisfies $\hat{\sigma}^*_{\hat{\lambda}}/\sigma_{\hat{\lambda}} \to 1$ in probability.

We can therefore use Wald tests and intervals for λ that are based on the approximation $(\hat{\lambda} - \lambda_0)/\hat{\sigma}_{\hat{\lambda}} \approx \mathsf{N}(0,1)$. The standard error of $\hat{\lambda}$ can also be estimated by

$$1 \Big/ \left\{\hat{g}^*\sqrt{-\ell''_{BC}(\hat{\lambda})}\right\}.$$

4.2. Parameters θ and β.

Assume Theorem 2 and 3 and also that $\max(H_{ii}) \to 0$. Then we have

$$\frac{a'(\hat{\theta} - \theta)}{\sqrt{a'\hat{\mathbf{V}}a}} \Rightarrow \mathsf{N}(0,1).$$

Confidence intervals and tests for $a'\theta$ can then be found. Testing $a'\theta = 0$ is equivalent to testing $a'\beta = 0$. If $a'\beta = 0$, then the asymptotic variance of $a'\hat{\theta}$ is

$$a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a + \frac{(a'v)^2 g^*}{4(g^{**})^2},$$

and the second term then provides a rather simple correction term to the variance given by standard linear regression. Notice that if $v = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\eta^{\#2} = 0$ (or even, of course, if merely $a'v = 0$), there is no variance inflation to worry about.

We now consider parameter β. If $\delta g^{1/2} \to \infty$ as $n \to \infty$, then $\hat{\phi}/\delta \to 0$ in probability. Confidence intervals for β may then be found by linearizing the term $(1 + \lambda_0\mu_o)^{\hat{\phi}/\delta}$ in (6), but the resulting normal approximations are complicated and will not behave well if the convergence of $\delta\sqrt{g} \to \infty$ is slow. Also, under the conditions of this paper, $\delta\sqrt{g}$ need not diverge, so that the linearization may not be valid. Finally, recall that confidence intervals based on $\hat{\beta} - \beta$ can be regarded as of little scientific use since they are derived from subtraction of incomparables. For these reasons, we do not consider intervals for β further.

4.3. Values of g for some special cases.

Some insight into the three terms in g can be gained by consideration of special cases. The K-sample analysis of variance problem can be cast in the linear model framework. Suppose we have K samples, where sample i is Y_{ij}, $j = 1, \ldots, n_i$ and $\sum n_i = n$. The Box–Cox model becomes $Y_{ij}(\lambda) = \mu + \alpha_i + \varepsilon_i$ where we impose the constraint $\sum n_i\alpha_i = 0$. Take $\beta_i = \alpha_i$ for $i = 1, \ldots, K-1$. The resulting design matrix is correctly centered. In this model, g reduces to

$3n/2 + \sum n_i \alpha_i^2$; the final term $(\eta^{\#2})'\mathbf{M}(\eta^{\#2})/4$ is 0. Similarly, in simple linear regression with equally spaced covariate values x_i, we would find that the term $(\eta^{\#2})'\mathbf{M}(\eta^{\#2})/4$ vanishes.

For other designs, it may be expected that all three terms in g will be important. Consider the situation in which the effects are large, in the sense that $\sum \eta_i^2$ is large compared to n. The term $(\eta^{\#2})'\mathbf{M}(\eta^{\#2})/4$ involves $\sum \eta_i^4$, which might be expected to be even larger than $\sum \eta_i^2$ which would make this last term in g dominant. This contrast between the K-sample design and more general designs has been discussed, for instance, by Hooper & Yang (1997).

5. TESTS OF NORMALITY

5.1. Weak convergence of the empirical process.

In Sections 2 and 3, we frequently obtained better results on the assumption that the errors are normally distributed. In this section, we consider the problem of testing this hypothesis. Again we consider a sequence of Box–Cox models indexed by n, suppressing the dependence of $\mathbf{X}$, θ, δ and p on n wherever possible. Let $\widehat{Y}_n(t)$ be the empirical process of the fitted residuals, namely,

$$\widehat{Y}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big[1\{\Phi(\hat{\varepsilon}_i) \le t\} - t\big].$$

We will strengthen the conditions (8) and (9) to

$$p^3 \log^2(p)/n \to 0 \tag{18}$$

and

$$\delta^2 p \to 0. \tag{19}$$

Then we have the following theorem.

THEOREM 4. *Consider a sequence of designs satisfying* (7), (10), (18), *and* (19). *Assume the errors have a standard normal distribution and*

$$\delta \log^{3/2}(n) \to 0.$$

Assume that $g/n \to \gamma \in [3/2, \infty)$ as $n \to \infty$. Then the empirical process $\widehat{Y}_n(t)$ converges weakly to a Gaussian process $Y(t)$ with zero mean and covariance function

$$\rho(s,t) = \min(s,t) - st - J_1(s)J_1(t) - \frac{1}{2}J_2(s)J_2(t) - \frac{1}{4\gamma}J_3(s)J_3(t),$$

where $J_1(t) = \Phi'\{\Phi^{-1}(t)\}$, $J_2(t) = \Phi^{-1}(t)J_1(t)$, $J_3(t) = [\{\Phi^{-1}(t)\}^2 - 1]J_1(t)$, Φ is the standard normal distribution function and $s,t \in [0,1]$.

The lengthy proof is outlined in Appendix B.

5.2. EDF tests of normality.

To test the hypothesis of normal errors when fitting model (2) to data, we use the well-known Cramér–von Mises family of EDF statistics. See, for example, Durbin (1973) and Stephens (1986). For each i, let $v_i = \Phi(\hat{\varepsilon}_i)$ and let the empirical distribution function of the v_i be

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1(v_i \le t), \quad 0 \le t \le 1.$$

The EDF statistics are based on the discrepancies between $\widehat{F}_n(t)$ and $F(t) \equiv t$ $(0 \leq t \leq 1)$, namely,

$$Q_n = n \int_0^1 \left\{\widehat{F}_n(t) - t\right\}^2 \psi(t)\, dt,$$

where $\psi(t) > 0$ is a suitable known weight function. As special cases, the Cramér–von Mises statistic W^2 is obtained when $\psi(t) \equiv 1$, and the Anderson–Darling statistic A^2 is obtained when $\psi(t) = 1/\{t(1-t)\}$. For a given data set $v_1, \dots, v_n$, with order statistics $v_{(1)} < \cdots < v_{(n)}$, W^2 and A^2 can be computed easily as

$$W^2 = \sum_{i=1}^{n} \left(v_{(i)} - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}, \tag{20}$$

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left\{ (2i-1) \log v_{(i)} + (2n+1-i) \log(1 - v_{(i)}) \right\}. \tag{21}$$

5.3. Calculation of test statistics.

To perform a goodness-of-fit test of $\mathcal{H}_0$: model (2) fits the data with $F_0 = \Phi$, the following steps are needed:

(a) Write the original model matrix in the form $\mathbf{X}^* = (\mathbf{1} \,|\, \mathbf{X}_1^*)$. Center matrix $\mathbf{X}_1^*$ into $\mathbf{X}_1$ by subtracting from each column the mean of that column and let $\mathbf{X} = (\mathbf{1} \,|\, \mathbf{X}_1)$. Let x_i denote the ith row of $\mathbf{X}_1$.

(b) Find $\hat{\lambda}$ by maximizing the log profile likelihood (3). We recommend that each observation Y_i be divided by the average $\sum Y_i/n$ before computing this profile to avoid some numerical difficulties.

(c) Compute $\hat{\mu}$, $\hat{\beta}$ and $\hat{\sigma}$ by regressing the transformed values $Y_i(\hat{\lambda}) = (Y_i^{\hat{\lambda}} - 1)/\hat{\lambda}$ on the design matrix $\mathbf{X}$. If $\mathbf{Y}(\hat{\lambda})$ has entries $Y_i(\hat{\lambda})$, then $\hat{\mu} = \sum_i Y_i(\hat{\lambda})/n$, $\hat{\beta} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}(\hat{\lambda})$ and $\hat{\sigma}^2 = \mathbf{Y}(\hat{\lambda})'\mathbf{M}\mathbf{Y}(\hat{\lambda})/(n-p-1)$, where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

(d) Compute standardized residuals $\hat{\varepsilon}_i = \left\{(Y_i^{\hat{\lambda}} - 1)/\hat{\lambda} - \hat{\mu} - x_i\hat{\beta}\right\}/\hat{\sigma}$.

(e) Compute $v_i = \Phi(\hat{\varepsilon}_i)$.

(f) Calculate W^2 or A^2 according to (20) or (21), respectively.

(g) Find $\hat{\eta}_i = x_i'\hat{\beta}/\hat{\sigma}$, and obtain an estimate $\hat{g}$ of g defined by

$$\hat{g} = \frac{3n}{2} + 2\sum_{i=1}^{n} \hat{\eta}_i^2 + \frac{1}{4}(\hat{\eta}^{\#2})'\mathbf{M}(\hat{\eta}^{\#2}).$$

(h) Enter Table 2 with the value of $n/\hat{g}$, and reject $\mathcal{H}_0$ at significance level α if the test statistic exceeds the corresponding upper α-percentile given in Table 2.

TABLE 2: Upper percentiles of the asymptotic distributions of W^2 and A^2 for testing Box–Cox transformations when $\delta = \lambda\sigma/(1+\lambda\mu) \to 0$ and $n \to \infty$.

Upper percentiles

					α			
Statistics	n/g	0.50	0.25	0.20	0.15	0.10	0.05	0.01
W^2	0	0.0508	0.0739	0.0812	0.0915	0.1036	0.1260	0.1787
	1/250	0.0508	0.0738	0.0810	0.0905	0.1031	0.1258	0.1785
	1/150	0.0508	0.0738	0.0810	0.0903	0.1031	0.1257	0.1783
	1/100	0.0507	0.0737	0.0809	0.0902	0.1030	0.1256	0.1781
	1/50	0.0506	0.0736	0.0807	0.0899	0.1028	0.1251	0.1773
	1/25	0.0504	0.0731	0.0802	0.0894	0.1022	0.1243	0.1759
	1/15	0.0501	0.0726	0.0796	0.0887	0.1011	0.1231	0.1740
	1/10	0.0498	0.0719	0.0787	0.0878	0.1002	0.1217	0.1716
	1/5	0.0487	0.0700	0.0766	0.0851	0.0970	0.1175	0.1649
	2/5	0.0463	0.0660	0.0721	0.0800	0.0909	0.1097	0.1530
	2/3	0.0428	0.0608	0.0663	0.0736	0.0836	0.1007	0.1406
A^2	0	0.3405	0.4702	0.5100	0.5607	0.6318	0.7530	1.0375
	1/250	0.3403	0.4697	0.5094	0.5601	0.6310	0.7520	1.0351
	1/150	0.3400	0.4693	0.5090	0.5596	0.6304	0.7512	1.0339
	1/100	0.3398	0.4689	0.5085	0.5590	0.6297	0.7504	1.0326
	1/50	0.3392	0.4677	0.0571	0.5574	0.6277	0.7476	1.0281
	1/25	0.3378	0.4653	0.5043	0.5541	0.6236	0.7422	1.0187
	1/15	0.3359	0.4620	0.5005	0.5496	0.6182	0.7351	1.0007
	1/10	0.3335	0.4578	0.4958	0.5441	0.6115	0.7262	0.9928
	1/5	0.3262	0.4454	0.4817	0.5277	0.5918	0.7004	0.9518
	2/5	0.3106	0.4202	0.4537	0.4958	0.5546	0.6537	0.8820
	2/3	0.2871	0.3880	0.4186	0.4575	0.5117	0.6035	0.8168

The entries in Table 2 are the upper tail percentiles of the asymptotic distributions of W^2 and A^2, respectively, as $n \to \infty$, and as $\delta \to 0$; see Subsection 5.4 below for the theory. Note that the percentiles corresponding to $n/g = 0$ are the upper tail percentiles for testing the goodness of fit of linear models *without* taking any Box–Cox transformation (see Stephens 1986, Section 4.8.5); the percentiles in the table differ increasingly from these values as n/g grows larger. The line $n/g = 2/3$ corresponds to models in which all the effects η_i are 0.

5.4. Asymptotic theory and calculation of Table 2.

The statistic W^2 is a continuous functional (on $D[0,1]$) and the weak convergence result of the last section guarantees that if the conditions of Section 4 hold and $g/n \to \gamma$, then W^2 converges in distribution to $\int_0^1 Y^2(t)\,dt$, where Y is the limiting Gaussian process of the previous section. The integrand may then be expanded as usual in orthonormal eigenfunctions of the operator $\rho(s,t)$ to show that the limiting distribution of W^2 is that of

$$\sum_{i=1}^{\infty} \xi_i Z_i^2, \tag{22}$$

where the Z_i are independent standard normals and the ξ_i are the eigenvalues of the integral equation $\int_0^1 \rho(s,t) f(t)\,dt = \xi f(s)$. We computed the ξ_i following the method of Stephens (1976).

We computed the largest 40 eigenvalues and replaced the sum in (22) by

$$\sum_{i=1}^{40} \xi_i Z_i^2 + \sum_{i=41}^{\infty} \xi_i, \tag{23}$$

computing the last term as $\int_0^1 \rho(t,t)\,dt - \sum_{i=1}^{40} \xi_i$. Table 2 gives critical points from the distribution of (23) calculated by the method of Imhof (1961), namely, numerical inversion of the characteristic function. In Section 7, we present some limited evidence of their accuracy. Our calculations for A^2 are not rigorous. The difficulty is that A^2 is not a continuous functional of the empirical process, involving as it does the unbounded weight function $\psi(t) = 1/\{t(1-t)\}$. The theory of the previous section permits us to prove that there is a sequence $\zeta_n \to 0$ such that

$$\int_{\zeta_n}^{1-\zeta_n} n\{F_n(t) - t\}^2 \psi(t)\,dt \;\Rightarrow\; \int_0^1 Y^2(t)\psi(t)\,dt.$$

The distribution of the limiting integral is of the same form as that for W^2 but with the kernel $\rho(s,t)$ replaced by $\rho(s,t)\psi(s)\psi(t)$. We conjecture that

$$\int_0^{\zeta_n} n\{F_n(t) - t\}^2 \psi(t)\,dt + \int_{1-\zeta_n}^{1} n\{F_n(t) - t\}^2 \psi(t)\,dt \to 0$$

in probability and hence that $A^2 \Rightarrow \int_0^1 Y^2(t)\psi(t)\,dt$ but are unable to complete the proof. This problem has occurred before; see, for instance, Durbin (1973). Nevertheless our Monte Carlo studies confirm this limit.

Linnet (1988) studied the use of the Anderson–Darling statistic A^2 and the Cramér-von Mises statistic W^2 to test for normality of the power transformed data for a sample Y with no covariates. Through simulation studies, Linnet concluded empirically that the null distributions of A^2 and W^2 do not depend on parameter values for transformation parameter λ, mean μ and variance σ^2. A table was provided for A^2 and W^2 for finite samples in which the asymptotic critical points were obtained by extrapolation.

6. EXAMPLES

Four examples are given below to illustrate the theory of the previous sections.

Example 1 (Gasoline data). We return to our opening example of gasoline mileage. We have $\hat\lambda = 1.46$ (standard error, SE =0.64), $\hat\theta = 0.344, \hat\delta = 0.097, \hat g/n = 5.1$. The standard error of $\hat\theta$, taking into account that both λ and σ have been estimated, is computed as after Theorem 2; the value is 0.047.

In (17), the term $vv'/(4g)$ gives the variance inflation due to estimation of λ; if this term is dropped, so that λ is assumed known and correct, the estimated SE of $\hat\theta$ is 0.041. Also, the term $\theta\theta'/(4n)$ gives the variance inflation due to estimating σ; if this term also is omitted (assuming σ to be known), the estimated SE of $\hat\theta$ becomes 0.034. We see that the effect of estimating the transformation parameter is not negligible even in the estimation of the SE of $\hat\theta$.

When estimating β, the variance inflation is huge in this example. If we transform the data using $\hat\lambda = 1.46$ and fit a straight line via ordinary least squares, treating λ as known and correct, we get a fitted slope $\hat\beta = 167.3$ with a standard error of 16.5. As noted in the Introduction, this is far smaller than the uncertainty of four orders of magnitude visible in the top right panel of Figure 1.

In Figure 1, $\log\hat\beta(\lambda)$ appears to depend linearly on λ. Turning to (6), we see that the leading term in $\log\hat\beta(\phi)$ is $\phi\log(1+\lambda_0\mu_0)/\delta$, which predicts linearity of the type shown. A similar observation holds for $\log\hat\sigma(\phi)$.

The goodness-of-fit statistics are $W^2 = 0.102$ and $A^2 = 0.667$, giving p-values of about 0.08 and 0.06, respectively; thus the assumption of normal errors is probably adequate.

There are grounds other than the quality of the normal assumption for believing that the model as posed provides a poor description of this data set. In particular, the model ignores an important serial correlation in the measurement of the amount of gasoline required to refill the tank, and it also ignores annual variations in gasoline consumption per kilometer due to changing weather. The purpose of using the model here is principally to illustrate the volatility of the estimates when slopes are not standardized by division by $\hat{\sigma}$.

Example 2 (Textile data). Table 4 of Box & Cox (1964) contains the result of a single replicate of a 3^3 factorial experiment. The response Y is the cycles to failures of worsted yarn. The three explanatory variables assume three different levels each; we code these three levels as -1, 0 and 1. See Box & Cox (1964) for details.

TABLE 3: EDF tests of fit for three main effect linear models for the textile data of Example 2.

Model	Parameter estimates				EDF statistic (p-value)	
	$\hat{\sigma}$	$\hat{\lambda}$	$\hat{g}/n$	$\hat{\delta}$	A^2	W^2
Y:	488.2	1	7.9	0.57	1.3523 (<0.01)	0.2364 (<0.01)
$Y(\hat{\lambda})$:	0.125	-0.060	262	-0.011	0.3372 (>0.50)	0.0495 (>0.50)
$\log Y$:	0.186	0	247	0	0.2480 (>0.50)	0.0323 (>0.50)

We fit three main effect linear models to the data using one slope for each factor. The first model uses Y directly. The second transforms Y according to the Box–Cox model (1) and the third uses the log transformation since the estimate ($\hat{\lambda} = -0.060$, SE = 0.065) of λ in the transformation is very close to 0. As can be seen from Table 3, the transformed models provide the better fits. Note that p-values for Y and $\log Y$ are based on asymptotic results treating λ as known and applying the theory of Stephens (1986, Section 4.8, Case 3).

Table 4 contains estimates $\hat{\theta}$ of the standardized slopes and the associated standard errors for the second model, i.e., for the Box–Cox model with $\hat{\lambda} = -0.060$. These standard errors are again calculated in three ways: SE_1 is calculated as though the estimates of σ and λ were correct and known, so that the effect of making estimates is ignored; SE_2 assumes λ is correct and known, but σ is estimated (thus ignoring the effect of estimating λ); and finally SE_3 is found assuming both σ and λ are estimated, using (17). In this example, there is little variance inflation due to estimation of λ (cf. SE_2 and SE_3), though estimation of σ makes an important difference (cf. SE_1 and SE_2).

TABLE 4: Parameter estimates $\hat{\theta}$ and standard errors for Example 2.

Covariate	Estimated coefficient $\hat{\theta}$	SE_1	SE_2	SE_3
Factor 1	4.56	0.236	0.664	0.664
Factor 2	-3.46	0.236	0.526	0.526
Factor 3	-2.15	0.236	0.376	0.376

Example 3 (Tree data). The tree data in *Minitab Student Handbook* (Ryan, Joiner & Ryan 1976, p. 278) are analyzed here. The heights (x_1), the diameters (x_2) at 4.5 feet above ground level and the volumes (Y) were measured for a sample of 31 black cherry trees in the Allegheny National Forest, Pennsylvania. The data were collected to determine an easy way of estimating the volume of a tree based on its height and diameter.

Again, three linear models are fitted to the data, using Y, $Y(\hat{\lambda})$ and $Y(1/3)$, where $1/3$ is chosen from the dimension of volume versus length.

Table 5 contains summary statistics and goodness-of-fit test results. The Box–Cox estimate $\hat{\lambda} = 0.307$ (with estimated standard error 0.087) is close to the estimate $1/3$ derived from dimensional considerations. All three models pass the EDF tests easily; as in Example 2 we compute p-values for Y and $Y(1/3)$ treating λ as fixed. For the model with $\hat{\lambda} = 0.307$, Table 6 contains estimates $\hat{\theta}$ of the standardized slope and the associated standard errors calculated as before in three ways. Again in this example, there is little variance inflation due to estimation of λ, though estimation of σ cannot be ignored.

TABLE 5: EDF tests of fit for three straight line models for the tree data of Example 3.

Model	Parameter estimates				EDF statistic (p-value)	
	$\hat{\sigma}$	$\hat{\lambda}$	$\hat{g}/n$	$\hat{\delta}$	A^2	W^2
Y:	3.882	1	126	0.13	0.2482 (>0.50)	0.0361 (> 0.50)
$Y(\hat{\lambda})$:	0.227	0.307	623	0.025	0.2925 (>0.50)	0.0450 (>0.50)
$Y(1/3)$:	0.249	$1/3$	625		0.2735 (>0.50)	0.0407 (>0.50)

TABLE 6: Parameter estimates $\hat{\theta}$ and standard errors for Example 3.

Covariate	Estimated coefficient, $\hat{\theta}$	SE$_1$	SE$_2$	SE$_3$
Diameter	1.824	0.068	0.241	0.247
Height	0.1763	0.0340	0.0400	0.0410

Example 4 (Biological data). In Table 2 of Box & Cox (1964), the entries are the survival times (unit is 10 hours) of animals in a 3×4 completely randomized factorial experiment. The factors are poison content, with three levels, and treatment, with four levels.

Following Box and Cox, three main effect models are fitted as in the two previous examples. Table 7 clearly shows that the two power transformations improve the model fit considerably. Box and Cox felt the inverse transform (model 3) was not only reasonable on scientific grounds, but was also supported by the data. Our tests of fit, with p-values calculated as in Examples 2 and 3, clearly support this view. The estimated standard error of $\hat{\lambda}$ is 0.189; thus $\lambda = -1$ is well within two standard errors of $\hat{\lambda} = -0.75$.

TABLE 7: EDF tests of fit for three main effect linear models for the biological data of Example 4.

Model	Parameter estimates			EDF statistic (p-value)	
	$\hat{\lambda}$	$\hat{g}/n$	$\hat{\delta}$	A^2	W^2
Y:	1	5.4	0.33	1.0373 (<0.05)	0.1572 (<0.05)
$Y(\hat{\lambda})$:	-0.75	15.7	-0.14	0.1974 (>0.50)	0.0281 (>0.50)
$Y(-1)$:	-1	16.2	-0.24	0.2861 (>0.50)	0.0387 (>0.50)

In these models, there are two parameters for the main effects of poison content and three for the treatment main effects. The actual value of $\hat{\theta}$ will depend on the particular choice of coding for the linear two factor additive model, and for this reason, we do not give a table of estimates of the components of θ and their standard errors. However, we have made calculations which suggest that variance inflation due to estimation of λ is again unimportant, but estimation of σ inflates standard errors for coefficients $\hat{\theta}$ by about 15% on average.

In Examples 2, 3 and 4 the estimate $\hat{\lambda}$ was found to be close to a "physically meaningful" value such as 0, 1/3 or -1. The scientist may well choose to take this value as providing a sensible transformation. We refer to this method for choosing λ as the "snap-to-the-grid" procedure by analogy with the jargon used in computer windowing systems which keep icons aligned. We are investigating how the expansions in Theorem 2 should be used to deal with this procedure.

7. MONTE CARLO STUDIES

The accuracy of the asymptotic points for the goodness-of-fit statistics, in Table 2, was investigated by a simulation study. Consider the tree data of Example 3 with $n = 31$ values; using the Box–Cox transformation with estimated λ the value of W^2 is 0.0450 and the p-value, given by the asymptotic points, is 0.590. The accuracy of this value was examined as follows. First, for the Box–Cox model, the estimates of the parameters were taken as the true values. A sample of size 31 was then simulated from this model, the Box–Cox transformation procedure was applied and the EDF statistics calculated. This was repeated 10,000 times so that estimates of the distributions of W^2 and A^2 were obtained. The fraction of W^2 values which exceeded 0.0450 gives the empirically derived p-value of the original data, for the statistic W^2; similarly for the statistic A^2. A similar study was made for the biological data of Example 4. Table 9 gives a comparison between the asymptotic p-values of the data and the empirical p-values. They are close enough to indicate that the asymptotic points in Table 2 can be safely used for samples of reasonable size (we suggest $n > 20$).

TABLE 9: Asymptotic and empirical p-values for the tree data (Example 3) and the biological data (Example 4).

Data		Asymptotic p-value	Empirical p-value
Tree	$\mathrm{P}(W^2 > 0.0450)$	0.5895	0.5833
	$\mathrm{P}(A^2 > 0.2925)$	0.6282	0.6100
Biological	$\mathrm{P}(W^2 > 0.0281)$	0.8700	0.8850
	$\mathrm{P}(A^2 > 0.1974)$	0.8848	0.9117

APPENDIX A: PROOFS FOR SECTION 3

It will be useful to review some notation. Recall that $U(\phi, \delta)$ is the vector with ith co-ordinate

$$U_i(\phi,\delta) = \frac{\{1+\delta(\eta_i+\varepsilon_i)\}^{1+\phi/\delta}-1}{(\phi+\delta)}.$$

Define

$$f(x,y) = \frac{(1+x)^{(x+y)/x}-1}{x+y}$$

and note that

$$U_i(\phi,\delta) = (\eta_i+\varepsilon_i) f\big\{\delta(\eta_i+\varepsilon_i), \phi(\eta_i+\varepsilon_i)\big\}.$$

The function f is analytic on $\{(x,y) : x > -1\}$ with

$$f(x,0) \equiv 1, \quad f(0,y) \equiv \frac{e^y-1}{y}, \quad f(x,-x) = \frac{\log(1+x)}{x}$$

and $f(0,0) = 1$. We write $f^{i,j}$ for $\partial^{i+j} f/\partial x^i \partial y^j$.

We let $u_i = U_i(0,0) = \eta_i + \varepsilon_i$. For vectors a and b, we let $a\#b$ be the vector with ith entry $a_i b_i$ and $a^{\#r}$ be the vector with ith entry a_i^r. The design matrix is $\mathbf{X} = (\mathbf{1} \mid \mathbf{X}_1)$, with $\mathbf{1}$ denoting a column vector with all entries equal to 1 and $\mathbf{X}_1$ satisfying $\mathbf{1}'\mathbf{X}_1 = 0$. We need the matrices $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$, $\mathbf{M} = \mathbf{I} - \mathbf{H}$, and $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)\mathbf{X}_1'$.

Our proof uses the Cauchy–Schwarz inequality many times in several forms. For ease of reference, we record two of those forms here. In the statements, $\mathbf{Q}$ is any of $\mathbf{M}$, $\mathbf{H}$ or $\mathbf{H}_1$, x_i is a row of $\mathbf{X}_1$, ψ is a p-vector and a and b are n-vectors.

$$\begin{aligned} |x_i\psi|^2 &\le x_i(\mathbf{X}_1'\mathbf{X}_1)^{-1}x_i'\psi'(\mathbf{X}_1'\mathbf{X}_1)\psi = H_{1,ii}\,\psi'(\mathbf{X}_1'\mathbf{X}_1)\psi && (24)\\ |a'\mathbf{Q}b|^2 &\le a'\mathbf{Q}ab'\mathbf{Q}b \le \|a\|^2\|b\|^2. && (25)\end{aligned}$$

Proof of Theorem 1. The theorem is an easy consequence of the following three lemmas.

LEMMA 1. *Under the conditions of Theorem 1,*

$$\ell'(0) = \sum_{i=1}^{n} \Omega_i + o_P\big(\sqrt{n}\big) \quad \text{and} \quad \ell'(0)/\sqrt{g^*} \Rightarrow \mathsf{N}(0,1).$$

LEMMA 2. *Under the conditions of Theorem 1, $-\ell''(0)/g^{**} \to 1$ in probability.*

LEMMA 3. *Under the conditions of Theorem 3, for each fixed $L > 0$*

$$\sup_{|\phi|<L/\sqrt{n}} \left\{ \left| \ell(\phi) - \phi\ell'(0) - \frac{\phi^2\ell''(0)}{2} \right| \right\} = o_P(1) \tag{26}$$

and

$$\sup_{|\phi|\le L/\sqrt{n}} \left\{ \frac{|\ell''(\phi) - \ell''(0)|}{n} \right\} = o_P(1). \tag{27}$$

Proof of Lemma 1. Through the proof, we will use the notation C for a constant which may change from use to use but does not depend on n, p, $\mathbf{X}$ or on any of the parameters. The profile score at $\phi = 0$ is given by

$$\ell'(0) = -n\frac{\varepsilon'\mathbf{M}W(\delta)}{\varepsilon'\mathbf{M}\varepsilon} + \frac{1}{\delta}\sum_{i=1}^{n}\log(1+\delta u_i),$$

where W has components $W_i(\delta) = u_i^2 f^{0,1}(\delta u_i, 0)$. Note that $W_i(0) = u_i^2/2$. Put $T_0 = \Omega_1 + \cdots + \Omega_n$. Write

$$\ell'(0) = T_0 + T_1 + T_2 + T_3 + T_4 + T_5,$$

where

$$\begin{aligned} T_1 &= \frac{1}{2n}\left(\sum_{i=1}^{n}\varepsilon_i\right)\left\{\sum_{i=1}^{n}(\varepsilon_i^2 - 1)\right\} + \varepsilon'\mathbf{H}(\eta\#\varepsilon) + \varepsilon'\mathbf{H}_1\varepsilon^{\#2},\\ T_2 &= \frac{1}{n}\varepsilon'\mathbf{1}\mathbf{1}'\{W(\delta) - W(0)\},\\ T_3 &= \varepsilon'\mathbf{H}_1\{W(\delta) - W(0)\},\\ T_4 &= \varepsilon'\mathbf{M}W(\delta)\left(1 - \frac{n}{\varepsilon'\mathbf{M}\varepsilon}\right)\end{aligned}$$

and

$$T_5 \;=\; \frac{1}{\delta}\sum_{i=1}^{n}\{\log(1+\delta u_i)-\delta u_i\}-\sum_{i=1}^{n}\varepsilon_i\{W_i(\delta)-W_i(0)\}.$$

The conditions (7) and (12) permit Taylor expansion of $W(\delta)$; we then make moment calculations to show that $T_i \to 0/\sqrt{n}$ in probability for $i = 1,\ldots,5$. All the arguments are similar; the case of T_5 is the most difficult and shows the importance of approximate normality for the ε_i, so we give details for this term only.

Let A_i denote the event $|\delta u_i| \le 1/2$ and $A = \cap A_i$. Write T_5 as $\sum T_{5,i}$. In view of (7) and (12), we have $\mathrm{P}(A) \to 1$ and $\mathrm{P}(T_5 = \sum T_{5,i}1_{A_i}) \to 1$. Consider first the case of normal errors. Integration by parts of $\varepsilon_i W_i(\delta)$ yields

$$\mathrm{E}\,(T_{5,i}1_{A_i}) = \frac{C_+}{\delta^2}\Phi'\left(\frac{-2\delta\eta+1}{\delta}\right)-\frac{C_-}{\delta^2}\Phi'\left(\frac{-2\delta\eta-1}{\delta}\right),$$

where Φ' is the standard normal density, $C_+ = 3\{\log(3/2)-1\}/2$, and $C_- = \{\log(1/2)-1\}/2$. Conditions (7) and (A1) then give $\mathrm{E}\,\bigl(\sum T_{5,i}1_{A_i}\bigr)/\sqrt{n} \to 0$. A Taylor expansion of $T_{5,i}$ in δ shows

$$|T_{5,i}1_{A_i}| \le \delta(C_1 + C_2|\eta_i|^3 + C_3|\varepsilon_i|^3)$$

for suitable constants C_i. Conditions (10) and (11) give $\mathrm{var}\bigl(\sum T_{5,i}1_{A_i}\bigr)/n \to 0$ so $T_5 = o_P(n^{1/2})$.

If the errors are not normal, we expand $T_{5,i}$ in a Taylor series in δ to terms in δ^{q-3} to get

$$T_5 = \sum_{j=1}^{q-3}\delta^j\sum_{i=1}^{n}T_{5,ij}+\sum_{i=1}^{n}R_i,$$

where $T_{5,ij}$ has the form $C_j\{u_i^{j+1}-\varepsilon_i u_i^{j+2}/(j+2)\}$. If the errors have the same first q moments as the normal distribution, then $\mathrm{E}\,(T_{5,ij}) = 0$. A variance calculation and conditions (10) and (A2.2) show that

$$\frac{1}{\sqrt{n}}\delta^j\sum_{i=1}^{n}T_{5,ij} \to 0$$

in probability for $j \le q-3$. As with normal errors,

$$\sum_{i=1}^{n}R_i = \sum_{i=1}^{n}R_i 1_{A_i},$$

except on an event whose probability tends to 0. On A_i, the remainder satisfies

$$|R_i| \le C\delta^{q-2}(|u_i|^{q+2}+|u_i|^{q+3}).$$

Conditions (7), (10), (A2.3) and calculation of the mean show $\sum_{i=1}^{n} R_i/\sqrt{n} \to 0$ in probability and therefore $T_5 = o_P\left(\sqrt{n}\,\right)$.

The first conclusion of Lemma 1 now follows. The proof of the lemma is finished by using conditions (10) and (11) to permit application of the central limit theorem to conclude that

$$\left(\frac{\varepsilon' r}{2}+\sum_{i=1}^{n}\eta_i\varepsilon_i^2+\frac{1}{2}\sum_{i=1}^{n}\varepsilon_i^3-\frac{3}{2}\sum_{i=1}^{n}\varepsilon_i\right)\Big/\sqrt{g^*} \;\Rightarrow\; \mathsf{N}(0,1).$$

Proof of Lemma 2. Write

$$-\ell''(0) = \frac{W(\delta)'\mathbf{M}W(\delta)+\varepsilon'\mathbf{M}W_2(\delta)}{S}-2\frac{\{\varepsilon'\mathbf{M}W(\delta)\}^2}{nS^2}, \tag{28}$$

where now

$$W_{2,i}(\delta) = u_i^3 f^{0,2}(0, \delta u_i), \quad W_{2,i}(0) = u_i^3/3,$$

and $S = \varepsilon'\mathbf{M}\varepsilon/n$. The last term in (28) is $o_P(n)$, from the proof of Lemma 1. Note that $S = 1 + o_P(1)$. Routine Taylor expansions and (25) (details are in Chen, Lockhart & Stephens 2001) show that

$$W(\delta)'\mathbf{M}W(\delta) = W(0)'\mathbf{M}W(0) + o_P(n) \quad \text{and} \quad \varepsilon'\mathbf{M}\{W_2(\delta) - W_2(0)\} = o_P(n).$$

Thus we find $-\ell''(0) = W(0)'\mathbf{M}W(0) + \varepsilon'\mathbf{M}W_2(0) + o_P(n)$. We need then only show

$$W(0)'\mathbf{M}W(0) = \frac{1}{4}\left\{(\eta^{\#2})'\mathbf{M}(\eta^{\#2}) + 4\sum_{i=1}^{n}\eta_i^2 + n(\mu_4 - 1)\right\} + o_P(n) \tag{29}$$

and

$$\varepsilon'\mathbf{M}W_2(0) = n\mu_4/3 + \sum_{i=1}^{n}\eta_i^2 + o_P(n). \tag{30}$$

Assertions (29) and (30) can be established by moment calculations. □

Proof of Lemma 3. Statement (26) follows by Taylor expansion from (27). In what follows, we write $U_\delta(\phi)$ for $U(\phi, \delta)$ and let superscript $'$ denote differentiation with respect to ϕ. Since

$$-\ell''(\phi) = \frac{U_\delta(\phi)'\mathbf{M}U_\delta''(\phi)}{U_\delta(\phi)'\mathbf{M}U_\delta(\phi)/n} + \frac{U_\delta'(\phi)'\mathbf{M}U_\delta'(\phi)}{U_\delta(\phi)'\mathbf{M}U_\delta(\phi)/n} - \frac{1}{n}\left\{\frac{U_\delta'(\phi)'\mathbf{M}U_\delta(\phi)}{U_\delta(\phi)'\mathbf{M}U_\delta(\phi)/n}\right\}^2,$$

the lemma follows easily from the following convergences:

$$\begin{aligned}
\sup_{|\phi|\le L/\sqrt{n}}\left\{\left|\frac{1}{n}U_\delta(\phi)'\mathbf{M}U_\delta(\phi) - 1\right|\right\} &= o_P(1)\\
\frac{1}{n^2}\sup_{|\phi|\le L/\sqrt{n}}\left\{\left|\left[U_\delta'(\phi)'\mathbf{M}U_\delta(\phi)\right]^2 - \left[U_\delta'(0)'\mathbf{M}U_\delta(0)\right]^2\right|\right\} &= o_P(1)\\
\frac{1}{n}\sup_{|\phi|\le L/\sqrt{n}}\left\{\left|U_\delta''(\phi)'\mathbf{M}U_\delta(\phi) - U_\delta''(0)'\mathbf{M}U_\delta(0)\right|\right\} &= o_P(1)\\
\frac{1}{n}\sup_{|\phi|\le L/\sqrt{n}}\left\{\left|U_\delta'(\phi)'\mathbf{M}U_\delta'(\phi) - U_\delta'(0)'\mathbf{M}U_\delta'(0)\right|\right\} &= o_P(1).
\end{aligned} \tag{31}$$

These assertions all have similar proofs; we do only (31). We use Taylor expansion and the elementary inequality, valid for any n-vectors a and b,

$$|a'\mathbf{M}a - b'\mathbf{M}b| \le (a-b)'\mathbf{M}(a-b) + 2\sqrt{\{(a-b)'\mathbf{M}(a-b)\}(b'\mathbf{M}b)}\,. \tag{32}$$

Let B_i be the event $L|u_i|/\sqrt{n} \le 1/2$ and $B = \cap B_i$. Use (10) and the fact $\mu_2 < \infty$ to see that $\mathbf{P}(B) \to 1$. On B, if $|\phi| \le L/\sqrt{n}$ we have $|U_{\delta,i}'(\phi) - U_{\delta,i}'(0)| \le C|u_i|^3/\sqrt{n}$. Hence

$$\sup_{|\phi|\le L/\sqrt{n}}\left[\{U_\delta'(\phi) - U_\delta'(0)\}'\mathbf{M}\{U_\delta'(\phi) - U_\delta'(0)\}\right] \le \frac{C}{n}\sum_{i=1}^{n}(\eta_i^6 + \varepsilon_i^6) = o_P(n). \tag{33}$$

Noting that $U_\delta'(0) = W(\delta)$ we see, as in the proof of Lemma 1, that

$$U_\delta'(0)'\mathbf{M}U_\delta'(0) = O_P(n).$$

Combining this with (33) and (32), we get (31).

Proof of Theorem 2. Recall that $(n-p-1)\tilde{\sigma}^2(\phi)/\sigma^2 = U(\phi,\delta)'\mathbf{M}U(\phi,\delta)$. A Taylor expansion and (25) establish

$$\sup_{|\phi|\le L/\sqrt{n}} \big\{\{U(\phi,\delta)-U(0,\delta)\}'\mathbf{M}\{U(\phi,\delta)-U(0,\delta)\}\big\} = O_P(1) = o_p\big(\sqrt{n}\,\big).$$

Expand $U_i(\phi,\delta)-U_i(0,\delta)$ as

$$\frac{\phi u_i^2}{2}+\frac{(c_{1,i}\phi^2+c_{2,i}\phi\delta)u_i^3}{6}, \tag{34}$$

where $c_{1,i}$ and $c_{2,1}$ are bounded on the event $A\cap B$. It follows, using $\mathbf{M}\eta=0$, that

$$U(0,\delta)'\mathbf{M}\{U(\phi,\delta)-U(0,\delta)\} = \phi\varepsilon'\mathbf{M}u^{\#2}+R.$$

Use Cauchy–Schwarz to see that

$$\sup_{|\phi|\le L/\sqrt{n}} |R| \le O_P(1)+\delta O_P\big(\sqrt{n}\,\big) = o_P\big(\sqrt{n}\,\big).$$

The mean of $\varepsilon'\mathbf{M}u^{\#2}$ is

$$\sum_{i=1}^n M_{ii}\eta_i = -\sum_{i=1}^n H_{ii}\eta_i;$$

its variance is $O(n)$. Since

$$\left|\sum_{i=1}^n H_{ii}\eta_i\right|^2 \le \left(\sum_{i=1}^n H_{ii}^2\right)\left(\sum_{i=1}^n \eta_i^2\right) \le p\sum_{i=1}^n \eta_i^2,$$

we see that

$$\varepsilon'\mathbf{M}u^{\#2} = O\big(\sqrt{pn}\,\big)+O_P\big(\sqrt{n}\,\big).$$

Thus

$$\sup_{|\phi|\le L/\sqrt{n}} |\phi\varepsilon'\mathbf{M}u^{\#2}| = O_P\big(\sqrt{p}\,\big) = o_P\big(\sqrt{n}\,\big).$$

Hence

$$\sup_{|\phi|\le L/\sqrt{n}}\left\{\left|\tilde{\sigma}^2(\phi)/\sigma^2-\frac{\varepsilon'\mathbf{M}\varepsilon}{n-p-1}\right|\right\} = o_P\big(1/\sqrt{n}\,\big),$$

from which (13) follows.

We now expand U as in(34) to yield, for any n-vector a,

$$a'\tilde{\beta}(\phi)/\sigma = a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\{\eta+\varepsilon+\tfrac{1}{2}\phi(\eta^{\#2}+2\eta\#\varepsilon+\varepsilon^{\#2})+\phi^2R_1+\phi\delta R_2\}, \tag{35}$$

where the remainder vectors R_1, R_2 have $|R_{j,i}|\le C(|\eta_i|^3+|\varepsilon_i|^3)$ for $j=1,2$. The Cauchy–Schwarz inequality implies that

$$|a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'R_j|^2 \le a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a\sum_{i=1}^n R_{j,i}^2.$$

Since $\sum_{i=1}^n(\eta_i^6+\varepsilon_i^6)=O_P(n)$, the remainder in (35) is

$$\sqrt{a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a}\,O_P\big(\sqrt{n}\,\big)\left\{O_P\left(\frac{1}{n}\right)+O_P\left(\frac{\delta}{\sqrt{n}}\right)\right\} = \sqrt{a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a}\,o_P(1).$$

The quantity $a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon^{\#2}$ has mean 0 and variance $\mu_4 a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a$. Hence

$$\phi a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon^{\#2} = \sqrt{a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a}\,O_P\big(1/\sqrt{n}\,\big).$$

The term $a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\eta\#\varepsilon)$ has mean 0 and variance $\sum w_i^2\eta_i^2$, where $w = a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. This variance is bounded above by

$$\max_{1\le i\le n}(\eta_i^2) \times \sum_{i=1}^n w_i^2 = o(n)\bigl\{a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a\bigr\}.$$

It follows that

$$\phi a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\eta\#\varepsilon) = \sqrt{a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a}\, o_P(1).$$

Now assemble the pieces to get

$$a'\tilde{\beta}(\phi)/\sigma = a'\theta + a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon + \tfrac{1}{2}\,\phi a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\eta^{\#2} + \sqrt{a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a}\, o_P(1).$$

Notice that the term $a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon$ has variance $a'(\mathbf{X}_1'\mathbf{X}_1)^{-1}a$, so that the last term is negligible relative to the other terms. This establishes (14).

The expansion (15) of $a'\hat{\theta}$ is an easy consequence of (13) and (14).

Derivation of equation (16). We have

$$\hat{\theta} - \theta = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\{U(\hat{\phi},\delta) - U(0,\delta)\} + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon$$

and

$$(\hat{\theta}-\theta)'\mathbf{X}_1'\mathbf{X}_1(\hat{\theta}-\theta) \le 2\bigl[\{U(\hat{\phi},\delta) - U(0,\delta)\}'\mathbf{H}_1\{U(\hat{\phi},\delta) - U(0,\delta)\} + \varepsilon'\mathbf{H}_1\varepsilon\bigr].$$

The second term on the right-hand side of this inequality has expectation p. The first is bounded by

$$C\hat{\phi}^2 m\left(\sum_{i=1}^n \eta_i^4 + \sum_{i=1}^n \varepsilon_i^4\right) = O_P(1).$$

Statement (16) follows.

Proof of Theorem 3. We will prove only the first statement as an illustration of the general tactics. Note first that

$$\sum_{i=1}^n \hat{\eta}_i^2 - \sum_{i=1}^n \eta_i^2 = \sum_{i=1}^n (\hat{\eta}_i - \eta_i)^2 + 2\sum_{i=1}^n \eta_i(\hat{\eta}_i - \eta_i).$$

The first term on the right is simply

$$(\hat{\theta}-\theta)'\mathbf{X}_1'\mathbf{X}_1(\hat{\theta}-\theta) = O_p(p) = o_P(n)$$

using Theorem 2. Apply Cauchy–Schwarz to $\left|\sum \eta_i(\hat{\eta}_i - \eta_i)\right|$ to get, on the event $A\cap B$, the bound

$$C\sqrt{p\left(\sum_{i=1}^n \eta_i^2\right)} = o_P(n).$$

Next, write

$$(\hat{\eta}^{\#2})'\mathbf{M}(\hat{\eta}^{\#2}) - (\eta^{\#2})'\mathbf{M}(\eta^{\#2}) = T_1 + 4T_2 + 4T_3 + 2T_4 + 4T_5$$

where

$$\begin{aligned}
T_1 &= \{(\hat{\eta}-\eta)^{\#2}\}'\mathbf{M}\{(\hat{\eta}-\eta)^{\#2}\},\\
T_2 &= \{\eta\#(\hat{\eta}-\eta)\}'\mathbf{M}\{\eta\#(\hat{\eta}-\eta)\},\\
T_3 &= \{\eta\#(\hat{\eta}-\eta)\}'\mathbf{M}\{(\hat{\eta}-\eta)^{\#2}\},\\
T_4 &= \{(\hat{\eta}-\eta)^{\#2}\}'\mathbf{M}(\eta^{\#2}),
\end{aligned}$$

and

$$T_5 \quad = \quad \{\eta\#(\hat{\eta}-\eta)\}'\mathbf{M}(\eta^{\#2}).$$

Apply the Cauchy–Schwarz inequality to conclude that

$$T_3^2 \le T_1 T_2, \quad T_4^2 \le CnT_1, \quad \text{and} \quad T_5^2 \le CnT_2.$$

This reduces the lemma to the claims $T_1 = o_P(n)$ and $T_2 = o_P(n)$. We give details for T_1.

Note that

$$T_1 \le \sum_{i=1}^{n}(\hat{\eta}_i - \eta_i)^4 = \sum_{i=1}^{n}\{x_i(\hat{\theta}-\theta)\}^4.$$

Apply (24) to get, on $A \cap B$,

$$|x_i(\hat{\theta}-\theta)|^2 \le H_{1,ii}(\hat{\theta}-\theta)'\mathbf{X}_1'\mathbf{X}_1(\hat{\theta}-\theta) \le LH_{1,ii}p.$$

Since $0 \le H_{1,ii} \le 1$, we find

$$T_1 \le Cp^2 \sum_{i=1}^{n} H_{1,ii} = Cp^3 = o(n),$$

and the proof is complete. □

APPENDIX B: PROOF OF THEOREM 4

The fitted residuals have the form $\hat{\varepsilon}_i = \tilde{\varepsilon}_i/(\tilde{\sigma}/\sigma)$, where

$$\tilde{\sigma}^2 = \sigma^2 U(\hat{\phi},\delta)'\mathbf{M}U(\hat{\phi},\delta)/(n-p-1)$$

and $\tilde{\varepsilon} = \mathbf{M}U(\hat{\phi},\delta)$. Write

$$\widehat{Y}_n(t) = \widetilde{Y}_n\{T_n(t)\} + \sqrt{n}\,\{T_n(t)-t\},$$

where

$$\widetilde{Y}_n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big[1\{\Phi(\tilde{\varepsilon}_i) \le t\} - t\big]$$

and $T_n(t) = \Phi\{\tilde{\sigma}\Phi^{-1}(t)/\sigma\}$. Our proof uses Taylor expansions of $\widetilde{Y}_n$ and of T_n. We prove the theorem via several lemmas whose proofs are postponed. All lemmas tacitly assume the hypotheses of Theorem 4. We work throughout on the event $A \cap B$, where A is defined in the proof of Lemma 1 and B is defined below (32).

First we study $\widetilde{Y}_n$ near 0 and 1.

LEMMA 4. *Let* $I_n = [n^{-3/4}, 1-n^{-3/4}]$. *Then* $\sup_{t\notin I_n}|\widetilde{Y}_n(t)| = o_P(1)$.

Finally, we expand $\widetilde{Y}_n$ when t is not too close to 0 or 1. Define

$$h(x,y) = \frac{\{1+(x+y)\}^{x/(x+y)}-1}{x}, \quad w_i = \Phi^{-1}(t) + \eta_i + x_i\psi + \nu$$

and

$$H_i(t,\phi,\psi,\nu) = \Phi\{w_i h(\delta w_i, \phi w_i) - \eta_i\}.$$

When $1 + (\phi+\delta)w_i \le 0$, extend the definition of H_i by putting $H_i = 0$ if $(\phi+\delta) > 0$ and $H_i = 1$ if $(\phi+\delta) < 0$. Then

$$\Phi(\tilde{\varepsilon}_i) \le t \;\Leftrightarrow\; \Phi(\varepsilon_i) \le H_i\{t,\hat{\phi},\hat{\psi},\overline{U}(\hat{\phi},\delta)\},$$

where

$$\hat{\psi} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\{U(\hat{\phi},\delta) - \eta\} \quad \text{and} \quad \overline{U}(\phi,\delta) = \frac{1}{n}\sum_{i=1}^{n} U_i(\phi,\delta).$$

Hence we can write

$$\widetilde{Y}_n(t) = Y_{1,n}(t) + Y_{2,n}(t) + Y_{3,n}(t),$$

where

$$\begin{aligned} Y_{1,n}(t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bigl[1\{\Phi(\varepsilon_i) \le t\} - t\bigr], \\ Y_{2,n}(t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bigl[H_i\{t,\hat{\phi},\hat{\psi},\bar{U}(\hat{\phi},\delta)\} - t\bigr], \\ Y_{3,n}(t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bigl[1\{\Phi(\tilde{\varepsilon}_i) \le t\} - 1\{\Phi(\varepsilon_i) \le t\} - H_i\{t,\hat{\phi},\hat{\psi},\overline{U}(\hat{\phi},\delta)\} + t\bigr] \end{aligned}$$

We expand $Y_{2,n}$ and show that $Y_{3,n}$ is negligible in the following two lemmas.

LEMMA 5. *The process $Y_{2,n}$ admits a Taylor expansion*

$$Y_{2,n}(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bigl[\varepsilon_i J_1(t) - \Omega_i J_3(t)/(2\hat{g}/n)\bigr] + R_{2,n}(t); \tag{36}$$

the remainder term $R_{2,n}$ satisfies $\sup_{t\in I_n} |R_{2,n}(t)| = o_P(1)$.

LEMMA 6. *We have* $\sup_{t\in I_n} |Y_{3,n}(t)| = o_P(1)$.

Put

$$\widetilde{Y}_n^*(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bigl[1\{\Phi(\varepsilon_i) \le t\} - t + \varepsilon_i J_1(t) - \Omega_i J_3(t)/(2\hat{g}/n)\bigr].$$

Since J_1 and J_3 vanish at 0 and 1, standard weak convergence methods lead to

$$\sup_{t\in I_n} |\widetilde{Y}_n^*(t)| = o_P(1).$$

With Lemmas 5 and 6, this shows

$$\sup_{0\le t\le 1} |\widetilde{Y}_n(t) - \widetilde{Y}_n^*(t)| = o_P(1).$$

It is then standard to check that $\widetilde{Y}_n$ converges weakly in $D[0,1]$ to a centered, continuous Gaussian process $\widetilde{Y}$ with covariance

$$\tilde{\rho}(s,t) = \min(s,t) - st - J_1(s)J_1(t) - \frac{1}{4\gamma}J_3(s)J_3(t).$$

Next, use (13) to show that

$$\sup_{0\le t\le 1} |R_n(t)| = o_P(1)$$

in the expansion

$$\sqrt{n}\,\{T_n(t) - t\} = \frac{1}{2\sqrt{n}}\sum(\varepsilon_i^2 - 1)J_2(t) + R_n(t).$$

Use this expansion, the weak convergence of $\widetilde{Y}_n$ and Lemma 4 to check that the process $\big(\widetilde{Y}_n, \sqrt{n}\,\{T_n(\,\cdot\,) - \cdot\,\}\big)$ is tight in $D[0,1] \times D[0,1]$. The Lindeberg central limit theorem shows the finite-dimensional distributions of that process to be asymptotically Gaussian with zero mean. The asymptotic covariance matrix between $\big(\widetilde{Y}_n(s), \sqrt{n}\,\{T_n(s) - s\}\big)$ and $\big(\widetilde{Y}_n(t), \sqrt{n}\,\{T_n(t) - t\}\big)$ is

$$\begin{pmatrix} \tilde{\rho}(s,t) & 0 \\ 0 & \frac{1}{2}\,J_2(s)J_2(t) \end{pmatrix}.$$

Hence $\big(\widetilde{Y}_n, \sqrt{n}\,\{T_n(\,\cdot\,) - \cdot\,\}\big)$ converges weakly in $D[0,1] \times D[0,1]$ to a continuous centered Gaussian process with this covariance function. The theorem now follows from the usual time transform argument (see Billingsley 1968, p. 145). □

Proof of Lemma 4. It evidently suffices to show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1\big\{|\tilde{\varepsilon}_i| > |\Phi^{-1}(n^{-3/4})|\big\} = o_P(1). \tag{37}$$

Write $\tilde{\varepsilon} = \mathbf{M}\{U(\hat{\phi},\delta) - U(0,\delta)\} + \varepsilon$ and bound (37) by the sum of the following two terms:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1\big\{|\varepsilon_i| > |\Phi^{-1}(n^{-5/8})|\big\} \tag{38}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1\Big[\big|\big[\mathbf{M}\{U(\hat{\phi},\delta) - U(0,\delta)\}\big]_i\big| > \big|\Phi^{-1}(n^{-3/4}) - \Phi^{-1}(n^{-5/8})\big|\Big]. \tag{39}$$

Since the ε_i are normal the expectation of (38) is $2/n^{1/8} \to 0$. For

$$\tau_n = |\Phi^{-1}(n^{-3/4}) - \Phi^{-1}(n^{-5/8})|,$$

we have $\tau_n \to \infty$ and

$$\begin{aligned} \sum_{i=1}^{n} 1\Big(\big|\big[\mathbf{M}\{U(\hat{\phi},\delta) - U(0,\delta)\}\big]_i\big| > \tau_n\Big) \quad &\le \quad \frac{\sum_{i=1}^{n}\big[\mathbf{M}\{U(\hat{\phi},\delta) - U(0,\delta)\}\big]_i^2}{\tau_n^2} \\ &= \quad \frac{\{U(\hat{\phi},\delta) - U(0,\delta)\}'\mathbf{M}\{U(\hat{\phi},\delta) - U(0,\delta)\}}{\tau_n^2}. \end{aligned}$$

That (39) converges to 0 can now be seen from the proof of (13) in Theorem 2. □

Proof of Lemma 5. The function h, in the definition of H_i, has a power series expansion over $\{(x,y) : |x+y| < 1\}$ and a Taylor expansion of the form $h(x,y) = 1 - y/2 + c_1(x)y^2 + c_2(x,y)xy$, where c_1 and c_2 are bounded over the set $|x| \le 1/2$, $|y| \le 1/2$, $|x+y| \le 1/2$. In order to apply this smoothness of h to carry out a Taylor expansion, we need to establish that the arguments of h are not too far from 0. We do this in the two following lemmas.

LEMMA 7. *For each $L < \infty$,*

$$\sup_{|\phi| \le L/\sqrt{n}} \big|\bar{U}_\delta(\phi) - \bar{\varepsilon} + \tfrac{1}{2}\phi\big(1 + \overline{\eta^2}\big)\big| = o_P(1).$$

Furthermore, for each $\zeta > 0$, there is a constant C such that

$$\mathrm{P}\big\{|\overline{U}_\delta(\hat{\phi})| \le C/\sqrt{n}\,\big\} \le \zeta.$$

Fix $L > 0$. Let $\mathcal{D}_1$ be the set of pairs (ψ,ν) such that $\psi'\mathbf{X}_1'\mathbf{X}_1\psi \le Lp$ and $\sqrt{n}\,|\nu| \le L$.

LEMMA 8. *We have*

$$\left(|\delta| + \frac{1}{\sqrt{n}}\right) \sup_{\mathcal{D}_1 \times I_n} \max_{1 \le i \le n} \left\{|\Phi^{-1}(t)| + |\eta_i| + |x_i\psi| + |\nu|\right\} \to 0.$$

In particular,

$$\left(|\delta| + \frac{1}{\sqrt{n}}\right) \sup_{t \in I_n} \max_{1 \le i \le n} \left\{|\Phi^{-1}(t)| + |\eta_i| + |x_i\hat{\psi}| + |\overline{U}_\delta(\hat{\phi})|\right\} = o_P(1).$$

The lemma permits us to place a uniform bound on all derivatives of h occurring in what follows. As in Appendix A, we use C as a generic constant which may change from place to place but does not depend on n or any parameter value.

Expand Φ about $\Phi^{-1}(t)$ to obtain, for a suitable t^*,

$$\begin{aligned}\sum_{i=1}^{n}\{H_i(t,\phi,\psi,\nu) - t\} &= \sum_{i=1}^{n}\left\{w_i h(\delta w_i, \phi w_i) - \eta_i - \Phi^{-1}(t)\right\} J_1(t) &(40)\\ &\quad + \frac{1}{2}\sum_{i=1}^{n}\left\{w_i h(\delta w_i, \phi w_i) - \eta_i - \Phi^{-1}(t)\right\}^2 J_2(t^*). &(41)\end{aligned}$$

Expand h to bound (41) by

$$C\left\{\sum_{i=1}^{n}(\nu + x_i\psi)^2 + \phi^2 \sum_{i=1}^{n} w_i^4\right\},$$

which is no more than

$$C\left\{n\nu^2 + \psi'\mathbf{X}_1'\mathbf{X}_1\psi + |\Phi^{-1}(t)|^4 + \nu^4 + \frac{1}{n}\sum_{i=1}^{n}(x_i\psi)^4 + \frac{1}{n}\sum_{i=1}^{n}\eta_i^4\right\}.$$

Let $\mathcal{D}$ denote the set of triples (ϕ,ψ,ν) such that $\sqrt{n}\,|\phi| < L$ and $(\psi,\nu) \in \mathcal{D}_1$. Using Lemma 7 and Theorem 2, we see that (10) implies that the supremum over $\mathcal{D} \times I_n$ of (41) is $o_P\left(\sqrt{n}\,\right)$.

Write the right-hand side of (40) as

$$\frac{1}{\sqrt{n}} J_1(t) \sum_{i=1}^{n}\left\{x_i\psi + \nu - \tfrac{1}{2}\phi w_i^2 + \phi^2 w_i^3 c_1(\phi w_i) + \phi\delta w_i^3 c_2(\delta w_i, \phi w_i)\right\}.$$

If

$$\frac{1}{\sqrt{n}} \sup_{\mathcal{D} \times I_n}\left\{\phi^2 \sum_{i=1}^{n} |w_i^3 c_1(\phi w_i)| + |\phi\delta| \sum_{i=1}^{n} |w_i^3 c_2(\delta w_i, \phi w_i)|\right\} = o(1) \tag{42}$$

and

$$\frac{1}{\sqrt{n}} \sup_{\mathcal{D} \times I_n}\left|\phi \sum_{i=1}^{n}\left[w_i^2 - \eta_i^2 - \{\Phi^{-1}(t)\}^2\right]\right| = o(1), \tag{43}$$

then the fact $\sum x_i\psi = 0$ will establish that, uniformly over ϕ, ψ, t, ν, the right-hand side of (40) is,

$$\frac{1}{\sqrt{n}} J_1(t)\left(n\nu - \frac{\phi}{2}\left[\sum_{i=1}^{n}\eta_i^2 + n\{\Phi^{-1}(t)\}^2\right]\right) + o(1).$$

Replace ν by $\overline{U}(\phi, \delta)$ and apply the expansion in Lemma 7 to get

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bigl[H_i\{t, \phi, \psi, \overline{U}(\phi, \delta)\} - t\bigr] \\ = \;& \frac{J_1(t)}{\sqrt{n}} \left(\sum_{i=1}^{n} \varepsilon_i + n\frac{\phi}{2}(1 + \overline{\eta^2}) - \frac{\phi}{2}\left[\sum_{i=1}^{n} \eta_i^2 + n\{\Phi^{-1}(t)\}^2\right] \right) + o_P(1) \\ = \;& \frac{J_1(t)}{\sqrt{n}} \left(\sum_{i=1}^{n} \varepsilon_i - \frac{n\phi}{2}\bigl[\{\Phi^{-1}(t)\}^2 - 1\bigr] \right) + o_P(1). \end{aligned}$$

Finally, replace ϕ by $\hat{\phi}$ and apply the expansion given in Theorem 1 to get (36).

It remains to establish the claims (42) and (43). To establish (42), we bound

$$\sum_{i=1}^{n} |w_i|^r \le C\left\{ n|\nu|^r + n\left|\Phi^{-1}\left(\frac{1}{n}\right)\right|^r + \sum_{i=1}^{n} |x_i\psi|^r + \sum_{i=1}^{n} |\eta_i|^r \right\}.$$

Use (24) to see that for $r > 2$, we have

$$\sum_{i=1}^{n} |x_i\psi|^r \le \sum_{i=1}^{n} H_{1,ii}^{r/2} (\psi'\mathbf{X}_1'\mathbf{X}_1\psi)^{r/2} \le p(\psi'\mathbf{X}_1'\mathbf{X}_1\psi)^{r/2}. \tag{44}$$

Apply Lemma 7, the fact that $|\Phi^{-1}(n^{-1})| = O\bigl\{\sqrt{\log(n)}\,\bigr\}$, (44) and (16) to show that

$$\sup\left\{ \sum_{i=1}^{n} |w_i|^3 \right\} = O\{n \log^{3/2}(n)\}.$$

Hence (42) is

$$O\left(\frac{1}{\sqrt{n}}\right)\left\{ O\left(\frac{1}{n}\right) + \delta\, O\left(\frac{1}{\sqrt{n}}\right) \right\} O\{n \log^{3/2}(n)\} = o(1).$$

To establish (43), we square out w_i^2 and get

$$\sum_{i=1}^{n} \bigl[w_i^2 - \eta_i^2 - \{\Phi^{-1}(t)\}^2\bigr] = \psi'\mathbf{X}_1'\mathbf{X}_1\psi + n\nu^2 + 2n\nu\Phi^{-1}(t) + 2\theta'\mathbf{X}_1'\mathbf{X}_1\psi.$$

Bound the first term using Theorem 3, the second using Lemma 7, and the third using Lemma 7 and $|\Phi^{-1}(t)| = O\bigl\{\sqrt{\log r(n)}\,\bigr\}$. For the last term, use the Cauchy–Schwarz inequality to get

$$\left| \sum_{i=1}^{n} \eta_i x_i \psi \right|^2 \le \sum_{i=1}^{n} \eta_i^2 \psi'\mathbf{X}_1'\mathbf{X}_1\psi \le Cnp.$$

Apply (10) to get (43). □

Proof of Lemma 6. We follow Loynes (1980), as adapted in Chen & Lockhart (2001); more details are in Chen, Lockhart & Stephens (2001). Our proof uses chaining (see Pollard 1984). Define

$$R_i(t, \phi, \psi, \nu) = 1\bigl\{\Phi(\varepsilon_i) \le H_i(t, \phi, \psi, \nu)\bigr\} - 1\{\Phi(\varepsilon_i) \le t\} - H_i(t, \phi, \psi, \nu) + t.$$

Let D denote the event that $\big(\hat\phi, \hat\psi, \overline{U}(\hat\phi, \delta)\big) \in \mathcal{D}$. For any $\zeta > 0$ if L is chosen sufficiently large we have, by Theorem 1, (13), (16) and Lemma 7 that $\mathrm{P}(D) > 1 - \zeta/2$. We work on the event $A \cap B \cap D$ with A and B as before; for all n sufficiently large, $\mathrm{P}(A \cap B \cap D) > 1 - \zeta$.

Define

$$R(t, \phi, \psi, \nu) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} R_i(t, \phi, \psi, \nu).$$

On the event D, we have

$$\sup_{t \in I_n} Y_{3,n}(t) \le R^- \equiv \sup_{t \in I_n, (\phi,\psi,\nu)\in\mathcal{D}} R(t, \phi, \psi, \nu).$$

We need only prove that $R^- = o_P(1)$ (and omit the parallel argument for the infimum).

For $\alpha > 0$, the covering number $M(\alpha)$ is the smallest integer m for which there exist $\psi_1, \ldots, \psi_m$ such that $\psi'(\mathbf{X}_1'\mathbf{X}_1)\psi \le Lp$ implies the existence of an integer k such that

$$(\psi - \psi_k)'(\mathbf{X}_1'\mathbf{X}_1)(\psi - \psi_k) < \alpha. \tag{45}$$

Let C_k denote the set of ψ in C satisfying (45). As in Chen & Lockhart (2001), we assume without loss $\psi_k \in C_k$ for each k. We take $\alpha = 1/(p \log n)$.

Let

$$N_t = \big[\sqrt{n}\,\log(n)\big], \quad N_\phi = \big[\sqrt{n}\,\log^{3/2}(n)\big], \quad \text{and} \quad N_\nu = \big[\sqrt{n}\,\big].$$

Put

$$z_j = j|\Phi^{-1}(n^{-3/4})|/N_t, \quad t_j = \Phi(z_j), \quad \phi_\ell = \ell L g^{-1/2}/N_\phi,$$

and $\nu_m = mL/\big(\sqrt{n}\,N_\nu\big)$. We study

$$R_{j,k,\ell,m} = \sup_{t_j \le t \le t_{j+1}, \psi \in C_k, \phi_\ell \le \phi \le \phi_{\ell+1}, \nu_m \le \nu \le \nu_{m+1}} R(t, \phi, \psi, \nu).$$

Note that

$$\mathrm{P}(R^- > \zeta) < \sum \mathrm{P}(R_{j,k,\ell,m} > \zeta),$$

where the sum is over $-N_t \le j \le N_t - 1$, $-N_\phi \le \ell \le N_\phi - 1$, $-N_\nu \le m \le N_\nu - 1$ and k for which C_k is not empty. There are no more than $8N_t N_\phi N_\nu M(\alpha)$ terms in this sum.

Fix j, k, ℓ, m. If $\psi \in C_k$, $\phi_\ell \le \phi \le \phi_{\ell+1}$, $\nu_m \le \nu \le \nu_{m+1}$ and $t_j \le t \le t_{j+1}$, then

$$\begin{aligned} R_i(t, \phi, \psi, \nu) &= 1\big\{\Phi(\varepsilon_i) \le H_i(t, \phi, \psi, \nu)\big\} - 1\{\Phi(\varepsilon_i) \le t\} - H_i(t, \phi, \psi, \nu) + t \\ &\le 1\big\{\Phi(\varepsilon_i) \le H_i(t_{j+1}, \phi_\ell, \psi_{ik}', \nu_{m+1})\big\} - 1\{\Phi(\varepsilon_i) \le t_j\} \\ &\quad - H_i(t_j, \phi_{\ell+1}, \psi_{ik}'', \nu_m) + t_{j+1}, \end{aligned} \tag{46}$$

where $\psi_{i,k}'$ maximizes $x_i\psi$ over C_k and $\psi_{i,k}''$ minimizes the same quantity. [Remark: the map $w \mapsto \Phi\{wh(\delta w, \phi w) - \eta\}$ is increasing for all w, δ, ϕ and η for which $1 + (\phi + \delta)w > 0$ while the map $\phi \mapsto \Phi\{wh(\delta w, \phi w) - \eta\}$ is decreasing in the same circumstances.] This eliminates t, ϕ, ψ, ν from the bound. Let

$$B_{i,j,k,\ell,m} = H_i(t_{j+1}, \phi_\ell, \psi_{ik}', \nu_{m+1}) - H_i(t_j, \phi_{\ell+1}, \psi_{ik}'', \nu_m) + t_{j+1} - t_j.$$

As in Loynes (1980), write the right-hand side of (46) as

$$\omega_{i,j,k,\ell,m}(X_{i,j,k,\ell,m} - p_{i,j,k,\ell,m}) + B_{i,j,k,\ell,m},$$

where $X_{i,j,k,\ell,m}$ is the Bernoulli variable $\big|1\{\Phi(\varepsilon_i) \le H_i(t, \phi, \psi, \nu)\} - 1\{\Phi(\varepsilon_i) \le t\}\big|$, $p_{i,j,k,\ell,m} = \mathrm{E}\,(X_{i,j,k,\ell,m})$ and $\omega_{i,j,k,\ell,m}$ is the sign of $H_i(t_{j+1}, \phi_{\ell+1}, \psi_{ik}', \nu_{m+1}) - t_j$.

LEMMA 9. *We have*

$$\frac{1}{\sqrt{n}} \max_{j,k,\ell,m} \sum_i B_{i,j,k,\ell,m} \to 0.$$

Put

$$\tau_{j,k,\ell,m} = \sum_{i=1}^{n} p_{i,j,k,\ell,m}(1 - p_{i,j,k,\ell,m}).$$

LEMMA 10. *There is a C with* $\max_{i,j,k,\ell,m} \tau_{j,k,\ell,m} \le \tau_b \equiv Cn^{1/2}\big(\log n + \sqrt{p}\,\big)$.

As in Chen & Lockhart (2001), use Lemmas 9 and 10 to show

$$\mathrm{P}(R^- > \zeta) < 8N_t N_\phi N_\nu M(\alpha) \exp\big\{-n\zeta^2/(8\tau_b + \sqrt{n}\,\zeta/6)\big\}.$$

Chen & Lockhart (2001) show $M(\alpha) \le \big(1 + \sqrt{2Lp/\alpha}\,\big)^p$. Take $\alpha = 1/(p \log n)$; for all large n, we have $M(\alpha) \le \{3Lp^2 \log(n)\}^{p/2}$. Combining these, we obtain

$$\begin{aligned} \log\{\mathrm{P}(R^- > \zeta)\} &\le C + \log(N_t N_\phi N_\nu) + \frac{p \log\{3Lp^2 \log(n)\}}{2} + \frac{-n\zeta^2}{8\tau_b + \sqrt{n}\,\zeta/6} \\ &= C + \frac{n}{\tau_b}\left[\frac{\tau_b \log(N_t N_\phi N_\nu)}{n} + \frac{\tau_b p \log\{3Lp^2 \log(n)\}}{2n} - \frac{\zeta^2}{8 + \sqrt{n}\,\zeta/(6\tau_b)}\right]. \end{aligned}$$

Our conditions imply $n/\tau_b \to \infty$. The quantity in square brackets converges to $-\zeta^2/8 < 0$. Hence $\log\{\mathrm{P}(R^- > \zeta)\} \to -\infty$, thereby completing the proof of the theorem. □

Proof of Lemma 7. The second statement follows from Theorem 1 and the first statement. Write

$$U_i(\phi, \delta) = u_i + \phi u_i^2/2 + R_i,$$

where on $A \cap B$ (defined as in Appendix A), we have

$$|R_i| \le C(\phi^2 + |\phi\delta|)(|\eta_i|^3 + |\varepsilon_i|^3).$$

Thus

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} |R_i| \le C\left(\frac{L^2}{n} + \frac{L\delta}{\sqrt{n}}\right) O_P(1) = o_P(1).$$

The lemma follows from

$$\sum_{i=1}^{n} u_i^2 = \left(\sum_{i=1}^{n} \eta_i^2 + 2\sum_{i=1}^{n} \eta_i \varepsilon_i + \sum_{i=1}^{n} \varepsilon_i^2\right) = \sum_{i=1}^{n} \eta_i^2 + O_P\big(\sqrt{n}\,\big) + n + O_P\big(\sqrt{n}\,\big). \qquad \square$$

Proof of Lemma 8. The lemma is an easy consequence of our conditions, Theorem 2, and Lemma 7 except for handling $|x_i\psi|$. We have $|x_i\psi|^2 \le \sum_{j=1}^{n} |x_j\psi|^2 = \psi' \mathbf{X}_1' \mathbf{X}_1 \psi \le Lp$ which, divided by n or multiplied by δ^2 converges to 0. □

Proof of Lemma 9. First

$$\max |t_{j+1} - t_j| \le C \max |z_{j+1} - z_j| \le C \frac{\log(n)}{n}.$$

Write $w_{i,1} = z_{j+1} + \nu_{m+1} + \eta_i + x_i\psi_k$ and $w_{i,2} = z_j + \nu_m + \eta_i + x_i\psi_k$. Let w_i' denote $w_{i,1}$ with ψ_k replaced by $\psi_{i,k}'$ and w_i'' denote $w_{i,2}$ with ψ_k replaced by $\psi_{i,k}''$. Consider now

$$\sum_{i=1}^{n} \big\{H_i(t_{j+1}, \phi_\ell, \psi_{ik}', \nu_{m+1}) - H_i(t_j, \phi_{\ell+1}, \psi_{ik}'', \nu_m)\big\} = S_1 + S_2 + S_3 + S_4,$$

where

$$\begin{aligned}
S_1 &= \sum_{i=1}^{n}\big[\Phi\big\{w_i' h(\delta w_i', \phi_\ell w_i') - \eta_i\big\} - \Phi\big\{w_{i,1} h(\delta w_{i,1}, \phi_\ell w_{i,1}) - \eta_i\big\}\big],\\
S_2 &= \sum_{i=1}^{n}\big[\Phi\big\{w_{i,1} h(\delta w_{i,1}, \phi_\ell w_{i,1}) - \eta_i\big\} - \Phi\big\{w_{i,2} h(\delta w_{i,2}, \phi_\ell w_{i,2}) - \eta_i\big\}\big],\\
S_3 &= \sum_{i=1}^{n}\big[\Phi\big\{w_{i,2} h(\delta w_{i,2}, \phi_\ell w_{i,2}) - \eta_i\big\} - \Phi\big\{w_{i,2} h(\delta w_{i,2}, \phi_{\ell+1} w_{i,2}) - \eta_i\big\}\big],\\
S_4 &= \sum_{i=1}^{n}\big[\Phi\big\{w_{i,2} h(\delta w_{i,2}, \phi_{\ell+1} w_{i,2}) - \eta_i\big\} - \Phi\big\{w_i'' h(\delta w_i'', \phi_{\ell+1} w_i'') - \eta_i\big\}\big].
\end{aligned}$$

We bound each S_i by Taylor expansion. For brevity, we do only S_1. The map $w \mapsto \Phi\{wh(\delta w, \phi w) - \eta\}$ has, by Lemma 8, a derivative which is bounded over all i, j, k, ℓ, and m. Hence

$$|S_1| \le C\sum_{i=1}^{n}|w_i' - w_{i,1}| \le C\sum_{i=1}^{n}|x_i(\psi_{ik}' - \psi_k)|.$$

From (24), we have $|x_i(\psi_{ik}' - \psi_k)|^2 \le H_{1,ii}(\psi_{ik}' - \psi_k)'\mathbf{X}_1'\mathbf{X}_1(\psi_{ik}' - \psi_k) \le H_{1,ii}\alpha$, so that

$$\frac{|S_1|}{\sqrt{n}} \le C\sqrt{\frac{\alpha}{n}}\sum_{i=1}^{n}\sqrt{H_{1,ii}} \le C\sqrt{\alpha p} = \frac{C}{\sqrt{\log n}}. \qquad \square$$

Proof of Lemma 10. We have

$$\tau_{j,k,\ell,m} = \sum_{i=1}^{n} p_{i,j,k,\ell,m}(1 - p_{i,j,k,\ell,m}) \le \sum_{i=1}^{n} p_{i,j,k,\ell,m},$$

where

$$\begin{aligned}
p_{i,j,k,\ell,m} &= \big|\Phi\{w_i' h(\delta w_i', \phi_k w_i') - \eta_i\} - \Phi(z_j)\big|\\
&\le \big|\Phi\{w_i' h(\delta w_i', \phi_k w_i') - \eta_i\} - \Phi(w_i' - \eta_i)\big| + \big|\Phi(w_i' - \eta_i) - \Phi(z_j)\big|.
\end{aligned}$$

The first term in the above bound is less than

$$\frac{C{w_i'}^2}{\sqrt{n}} \le \frac{C}{\sqrt{n}}\left\{z_{j+1}^2 + (x_i\psi_{ik})^2 + \nu_{m+1}^2 + \eta_i^2\right\}.$$

The second term is bounded above by $C\big(|z_{j+1} - z_j| + |x_i\psi_{ik}| + |\nu_{m+1}|\big)$. Sum these bounds over i. Apply Lemma 7 and the inequalities $|z_j| \le C\sqrt{\log(n)}$ and $|z_{j+1} - z_j| \le C\sqrt{\log(n)}\,/N_t$ to get

$$\tau_{j,k,\ell,m} \le C\left\{\sqrt{n}\,\log(n) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i\psi_{ik})^2 + \sum_{i=1}^{n}|x_i\psi_{ik}|\right\}.$$

Now $|x_i\psi_{ik}| \le |x_i(\psi_{ik} - \psi_k)| + |x_i\psi_k|$ and $(x_i\psi_{ik})^2 \le 2\big[\{x_i(\psi_{ik} - \psi_k)\}^2 + (x_i\psi_k)^2\big]$. Use (24) and (45) to see that $\sum_{i=1}^{n}\{x_i(\psi_{ik} - \psi_k)\}^2 \le p\alpha = 1/\log(n)$ and

$$\left\{\sum_{i=1}^{n}|x_i(\psi_{ik} - \psi_k)|\right\}^2 \le n\sum_{i=1}^{n}\{x_i(\psi_{ik} - \psi_k)\}^2 \le \frac{n}{\log(n)}.$$

We have $\sum_{i=1}^{n}(x_i\psi_k)^2 = \psi_k'\mathbf{X}_1'\mathbf{X}_1\psi_k \le Lp$ and, by the Cauchy–Schwarz inequality,

$$\left(\sum_{i=1}^{n}|x_i\psi_k|\right)^2 \le n\sum_{i=1}^{n}(x_i\psi_k)^2 \le Lnp.$$

Hence $\sum_{i=1}^n |x_i \psi_{ik}| \le \sqrt{Lnp} + \sqrt{n/\log(n)}$ and $\sum_{i=1}^n (x_i\psi_{ik})^2 \le 2\{Lp + 1/\log(n)\}$ giving

$$\tau_{j,k,\ell,m} \le C\{\sqrt{n}\,\log n + \sqrt{np}\}. \qquad \square$$

ACKNOWLEDGEMENTS

The authors are grateful to the former Editor, Professor Christian Genest, to the Associate Editor, and to several referees, whose comments improved the work greatly. Partial funding in support of this work was provided by the Natural 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.

P. Billingsley (1968). *Convergence of Probability Measures*. Wiley, New York.

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.

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.

G. Chen & R. A. Lockhart (2001). Weak convergence of the empirical process of residuals in linear models with many parameters. *The Annals of Statistics*, 29, 748–762.

G. Chen, R. A. Lockhart & M. A. Stephens (2001). *Large Sample Theory for Box–Cox Transformations in Linear Models*. Technical Report, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada.

D. R. Cox & N. M. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society Series B*, 49, 1–39.

K. A. Doksum & C.-W. Wong (1983). Statistical tests based on transformed data. *Journal of the American Statistical Association*, 78, 411–417.

N. R. Draper & D. R. Cox (1969). On distributions and their transformation to normality. *Journal of the Royal Statistical Society Series B*, 31, 472–476.

J. Durbin (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*. Regional Conference Series in Applied Mathematics, 9. SIAM, PA.

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 (with discussion). *Journal of the American Statistical Association*, 79, 302–309.

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

J. P. Imhof (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48, 419–426.

K.-C. Li & N. Duan (1989). Regression analysis under link violation. *The Annals of Statistics*, 17, 1009–1052.

K. Linnet (1988). Testing normality of transformed data. *Applied Statistics*, 37, 180–186.

R. M. Loynes (1980). The empirical distribution function of residuals from generalised regression. *The Annals of Statistics*, 8, 285–298.

D. Pollard (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

T. Ryan, B. Joiner & B. Ryan (1976). *Minitab Student Handbook*. Duxbury Press, North Scituate, MA.

M. A. Stephens (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *The Annals of Statistics*, 4, 357–369.

M. A. Stephens (1986). Tests based on EDF statistics. In *Goodness-of-Fit Techniques* (R. B. D'Agostino &

M. A. Stephens, eds.), Marcel Dekker, New York, pp. 97–193.

J. M. G. Taylor (1985a). Measures of location of skew distributions obtained through Box–Cox transformations. *Journal of the American Statistical Association*, 80, 427–432.

J. M. G. Taylor (1985b). Power transformations to symmetry. *Biometrika*, 72, 145–152.

J. M. G. Taylor (1986). The retransformed mean after a fitted power transformation. *Journal of the American Statistical Association*, 81, 114–118.

Received 26 February 2000
Accepted 30 June 2001

Gemai CHEN: gchen@math.ucalgary.ca
Department of Mathematics and Statistics
University of Calgary, Calgary, Alberta, Canada T2N 1N4

Richard A. LOCKHART: lockhart@stat.sfu.ca
Michael A. STEPHENS: stephens@stat.sfu.ca
Department of Statistics and Actuarial Science
Simon Fraser University, Burnaby
British Columbia, Canada V5A 1S6

Discussion[1]

Comment 1: Karim Maher ABADIR

I would like to congratulate the authors for a very well-written paper on an important subject. I have a few queries, mostly relating to interpretation, which I hope may lead to further insights into their new parameterizations and expansions.

The case of $\delta \to 0$ can lead to difficulties which may not be present in other parameterizations. The limit $\delta \to 0$ can arise from two observationally equivalent models, the logarithmic model or the small-σ asymptotics of Kadane, Bickel and Doksum. Should an investigator find $\hat{\delta} \approx 0$, what is the next step and how can the Type I error of inference be bounded if a sequential procedure is adopted?

The current parameterization can lead to another type of identification issue, namely the singularity of the information matrix (with respect to ϕ and θ) at a particular point in the parameter space; see, e.g., Davies (1977) and the subsequent literature, and Rothenberg (1971), Catchpole & Morgan (1997). How relevant would this problem be for typical hypotheses of interest? The joint (as opposed to marginal) expansions of $\hat{\phi}$ and $\hat{\theta}$ may provide an answer to this question.

Wald tests or intervals are not invariant to nonlinear transformations of the parameter space. Dagenais & Dufour (1991) have illustrated the problems that arise from this in a related (but different) Box–Cox model, and Critchley, Marriott & Salmon (1996) have provided differential-geometric insight into this problem. The use of such intervals for λ in Section 4.1 would give results that are different from the ones obtained by using intervals for ϕ than inferring the corresponding ones for the parameter of interest λ. Given the invariance of likelihoods, wouldn't it be preferable to use other types of intervals based on $\ell(\hat{\phi})$ or Z?

An interesting extension of the plots in Figure 1 would be to include $\lambda \to 0$. Would the β/σ parameterization still be the preferred one, or would a more general form emerge? Box–Cox transformations are a special case of hypergeometric functions, which can be used to estimate more general nonlinear transformations of the dependent variate; see Abadir (1999) and Lawford (2001). It would be interesting to see how far the intuition gained from this paper, on stabilizing parameterizations, can be extended to the more general case.

Karim Maher ABADIR: kma4@york.ac.uk
Department of Mathematics and Department of Economics
University of York, Heslington
York YO10 5DD, England, UK

[1] These comments reflect the discussion of the paper which occurred after its public reading at the Centre de recherches mathématiques de Montréal on March 22, 2002. / Ces commentaires reflètent les débats qui ont suivi la présentation de cet article au Centre de recherches mathématiques de Montréal le 22 mars 2002.

Comment 2: Nancy M. REID

The authors are to be congratulated for the formidable asymptotic analysis presented here and for the original and careful attention to testing for goodness of fit.

The Box–Cox model is very useful for the theory of statistics, as a moderately anomalous model in the sense that blind application of conventional theory leads to absurd results. Early attempts to fix this seemed somewhat ad hoc, but there are now several different arguments that lead to essentially the same conclusion, and in this light it seems very clear that the regression parameter β has no meaningful interpretation in this model. The common conclusion is that the meaningful parameter is $\theta = \beta/\sigma$; the most incisive (and intimidating) of the arguments is provided in recent work by McCullagh, the simplest is the development of parameters orthogonal to λ in Cox & Reid (1987).

Cox & Reid (1987) suggested using orthogonal parameters to construct an adjusted profile likelihood, one motivation being that the resulting confidence intervals based on the adjusted likelihood would allow for the fact that the regression parameters had been estimated. With large numbers of nuisance parameters, this adjustment is relatively more important. In the notation of Chen, Lockhart & Stephens, the adjusted log-profile likelihood is

$$\ell_{CR}(\lambda) = -\left(\frac{n-p}{2}\right)\log Y(\lambda)'(I-H)Y(\lambda) + \lambda\Sigma\log Y_i + (n-p)\log\lambda - p\log\{1'Y(\lambda)\},$$

whereas Chen, Lockhart and Stephens use the log-profile likelihood

$$\ell_{BC}(\lambda) = -\left(\frac{n}{2}\right)\log Y(\lambda)'(I-H)Y(\lambda) + \lambda\Sigma\log Y_i.$$

Presumably the most important difference between these two expressions is the degrees of freedom adjustment to the residual sums of squares. It would be interesting to know if this adjustment goes some way to making estimation of λ “more nearly consistent”.

One of the attractions of the Box–Cox model is that it provides a mathematical formulation of a strategy that seems clearly useful and is much used in applied work. In the context of applications, the strategy that the authors call “snap-to-the-grid” seems the most sensible, and likely the most widely used. In this light, the observation in the examples that there is very little variance inflation due to the estimation of λ seems potentially very important, and some further investigation of this would be very interesting. The estimation of the variance relies on sample moments up to order six, and I thought it was nearly axiomatic in applied statistics that these estimates, while consistent, are so variable that they are almost useless. As a final comment, if snap-to-the-grid turns out to be theoretically (nearly) correct, as well as sensible, then it is less clear whether tests of fit for the full model are needed, instead of simply using the usual diagnostics for the transformed data.

Nancy M. REID: reid@utstat.utoronto.edu
Department of Statistics, University of Toronto
100 Saint George Street, Toronto, Ontario, Canada M5S 3G3

Comment 3: Peter McCULLAGH

Roughly 25 years ago, when I first read the Box–Cox paper as a graduate student, I tried to replicate numerically the analysis of one of their data sets to convince myself that I understood exactly what they were saying. Naturally, I proceeded to substitute the value $\hat{\lambda}$ and to analyse the data as if that were the true value. When that failed to reproduce the values reported in the paper, I checked my calculations but I could find no error. Fortunately I was at Imperial College, so I knew who to ask and I did so at the next opportunity. Following that brief lesson and a closer reading of the paper, I can say with some assurance that the Box–Cox paper does not contain a recommendation, or even a suggestion, "that $\hat{\lambda}$ be treated as the true value for the purpose of further analysis".

The emphasis on the family of power transformations in applied statistics deserves a deeper explanation than convenience and monotonicity. The following argument, based on consistency under aggregation, shows that the power transformation is inevitable. Consider a field experiment in which each plot is two meters square, $Y(u)$ is the yield on plot u and $\mu(u) = \mathrm{E}\,\{Y(u)\}$ is the vector of mean values. It is natural to consider a linear model of the form $g(\mu) \in \mathbf{M}$ in which the link function $g\colon (0,\infty) \to \mathbb{R}$ acts componentwise, and $\mathbf{M} \subset \mathbb{R}^n$ is a linear subspace containing the subspace 1 of constant functions. For example, $\mathbf{M}$ might be the subspace consisting of functions of the form $x(u) = \alpha + \beta_{v(u)} + \gamma_{t(u)}$, where $v(u)$ is the variety and $t(u)$ is the treatment in plot u. Thus, the systematic component of the model is

$$g(\mu_{vt}) = \alpha + \beta_v + \gamma_t.$$

Consider also a second experiment of the same type in which the plots are two by four meters. The same reasoning leads us to consider a model

$$g(\mu'_{vt}) = \alpha' + \beta'_v + \gamma'_t$$

in which $\mu', \alpha', \beta', \gamma'$ are the parameters relevant to the larger plots. Additivity of yields implies that $\mu'_{vt} = 2\mu_{vt}$, which leads naturally to the compatibility condition

$$g\bigl\{\mu_{vt}(A)\bigr\} = g(A\,\mu_{vt}) = \alpha(A) + \beta_v(A) + \gamma_t(A)$$

for plots of area A. In other words, $g(x) \in \mathbf{M}$ implies $g(Ax) \in \mathbf{M}$ for all $A > 0$ and for each subspace $\mathbf{M}$ containing 1. It is easy to see that this condition is satisfied by all transformations of the form $g(x) = ax^\lambda + b$ or $a\log(x) + b$, and no others. The induced transformation on the parameters is $\beta_v(A) = A^\lambda \beta_v(1)$ for $\lambda \neq 0$, which makes it clear how inference is possible for treatment effects on plots of arbitrary size.

In the Box–Cox version, the transformation is applied to the random variable rather than to the mean value, which is a major complication. Thus, aggregation is a linear transformation $T\colon \mathbb{R}^n \to \mathbb{R}^m$, applied to the untransformed yields on replicate plots only. In general, $Y^\lambda \sim \mathsf{N}_n(\mathbf{M}_n, \sigma^2 I_n)$ is incompatible with the aggregated version $(TY)^\lambda \sim \mathsf{N}_m(\mathbf{M}_m, \sigma'^2 I_m)$, where $\mathbf{M}_n \subset \mathbb{R}^n$ and $\mathbf{M}_m \subset \mathbb{R}^m$ are the corresponding model subspaces. In practice, if σ^2 is small, the model is approximately compatible under aggregation.

Chen, Lockhart and Stephens describe the ratio parameter $\phi = \beta/\sigma$ in the Box–Cox model as "physically meaningful". As a discussant, it is hard to disagree with them, for they do not provide a definition of the term. I do not want to suggest that they are the first to use this term without definition: it may even occur in the Box–Cox paper! By implication, neither β nor σ is physically meaningful. But what about the pair (β, σ) or the pair (β, λ), or the scalar $\lambda^{-1}\log\beta$? How are we to tell which functions are physically meaningful and which are not?

The discussion of physical units in Section 1.2 suggests to me that the authors' usage of the term coincides with the notion in algebra of a natural transformation, or what I call a natural sub-parameter (McCullagh 2002). Many statistical models have the property that they are closed under a group, or semi-group, of transformations. I have in mind a simple group action such as a change in the response scale $y \mapsto ay$ from, say, liters into gallons or miles into kilometers, so that the group is the set of positive scalar multiples $a > 0$. There may also be a second group acting on the covariate scale, but I will ignore this for simplicity in the present discussion. Invariably there is an induced action $\theta \mapsto a^*\theta$ on the parameter space. In the case of the Box–Cox model, it is convenient to write

$$Y \sim \mathsf{N}(\mu, \sigma^2, \lambda),$$

for each component, meaning that $Y^\lambda \sim \mathsf{N}(\mu, \sigma^2)$ in the conventional notation. If there is concern over negative values, Y^λ may be interpreted as $\pm|Y|^\lambda$ maintaining the sign of Y. This is different from the authors' parameterization, but it is more convenient for studying the induced group action. Evidently,

$$Y^\lambda \sim \mathsf{N}(\mu, \sigma^2) \ \Rightarrow\ (aY)^\lambda \sim \mathsf{N}(a^\lambda\mu, a^{2\lambda}\sigma^2),$$

or $aY \sim \mathsf{N}(a^\lambda\mu, a^{2\lambda}\sigma^2, \lambda)$, so that the induced action on the parameter space is

$$a^*\colon (\mu, \sigma^2, \lambda) \mapsto (a^\lambda\mu, a^{2\lambda}\sigma^2, \lambda),$$

which is a homomorphism of the group by transformations $\Theta \to \Theta$ on the parameter space.

A sub-parameter $\phi = g(\theta)$ is called natural if the value of ϕ relevant to the transformed response aY, i.e., $g(a^*\theta)$, can be determined from the pair (a, ϕ). This definition can be stated more accurately in terms of commutative diagrams, but the preceding version suffices for present purposes, and helps to explain why the concept might be relevant. If we are told the value of ϕ in liters, can we determine the corresponding value in gallons? If ϕ is "physically meaningful", the answer must surely be yes, and that is how I propose to interpret the term.

In the case of the Box–Cox model with the indicated group of scalar multiples acting on the response, the parameter μ is not natural because the value $a^\lambda\mu$ on the transformed scale cannot be determined from the pair (a, μ) alone, as the definition requires. However, λ, $\lambda^{-1}\log\mu$ and μ/σ are natural, as are the pairs (μ, λ) and (σ^2, λ), and also ratios β_r/β_s of regression coefficients in the case of a multiple regression model $\mu = X\beta$. It should be self-evident that inference is possible only for natural sub-parameters, in which case the complications discussed by Bickel and Doksum do not seem to arise. But at first sight, it would seem that $\lambda^{-1}\log\mu$, or $\mu^{1/\lambda}$, might be a more useful parameter than μ/σ. Is there a compelling reason to prefer one over the other?

In view of these examples, I ask the authors if their concept of the term "physically meaningful" is the same as that of a natural sub-parameter? If not, what is the distinction? If they are equivalent, the concept is purely algebraic. Any connection with statistical variability, stability, consistency, large samples, or even epistemology, would appear to be largely fortuitous.

Peter McCULLAGH: pmcc@galton.uchicago.edu
Department of Statistics, University of Chicago
5734 South University Avenue, Chicago, IL 60637, USA

Comment 4: Peter J. BICKEL

Chen, Lockhart and Stephens return to the problem of analyzing the behaviour of parameters defined on the Box–Cox (1969) transformation models some twenty years after Bickel & Doksum (1981) initiated a flurry of activities in the early 1980s. The same model is the subject of an example in the paper "What is a statistical model?" by McCullagh (2002) which is contemporaneous with this one.

Chen, Lockhart & Stephens argue that the appropriate asymptotics for this problem is as the parameter $\delta = \lambda\sigma/(1 + \lambda\mu) \to 0$ and $n \to \infty$ and that the relevant parameter to estimate is $\theta = \beta/\sigma$. They argue on the grounds that this asymptotics is stable in that the limit is a continuous function of θ not exhibiting a discontinuity at $\theta = 0$, whereas the asymptotics of Bickel and Doksum letting $\sigma \to 0$ keeping β fixed does exhibit such a discontinuity.

Stability is a desirable property and as we stressed also, there is no need to limit oneself to the classical $n \to \infty$ asymptotics. The interpretation of $\delta \to 0$ is a little difficult, however. With σ tending to 0 at a known rate, we could at least state what the order of magnitude of the errors in relation to n should be. Also, asymptotic theories requiring errors which have the first six moments the same as the Gaussian leave me somewhat uncomfortable.

Having said this, I am entirely in favour of considering θ as the relevant parameter. This parameter, mentioned in connection with a number of people by Chen, Lockhart and Stephens, arises in an entirely natural way if one extends the Box–Cox model to the natural semiparametric model

$$a(Y) = \beta X + \varepsilon,$$

where a is an unknown monotone transformation rather than just a power and ε independent of X has $\mathsf{N}(\mu, \sigma^2)$ distribution. It is immediate that under these assumptions, β is not identifiable but β/σ is. As expected, a is identifiable up to a linear transformation. Methods for fitting this model are suggested in Bickel & Ritov (1997), among others.

Incidentally consideration of $\theta = \beta/\sigma$ deviates from (alleviates?) the need for specifying knowledge of the rate at which $\sigma \to 0$ in our asymptotics.

I'd like to make a comment about the tests for normality of the distribution of ε. In a paper recently submitted to *The Annals of Statistics*, Ritov, Ryden and I argue that in the context of the usual $n \to \infty$ asymptotics, we should tailor tests of goodness of fit to have power in directions which we deem are important and set critical values by simulation using the parametric bootstrap for instance. This may also be feasible in this case, but it is not immediately clear how the consistency of the bootstrap would be affected by $\delta \to 0$.

Chen, Lockhart and Stephens point out that under their asymptotic theory the interpretable parameter β/σ is not subject to variance inflation for $\beta = 0$ or various symmetric designs. It would be interesting to know to what extent this conclusion remains valid in the more general transformation model (1).

In conclusion, I enjoyed the delicate asymptotics of this approach, though I am not entirely persuaded of its applicability.

Peter J. BICKEL: bickel@stat.berkeley.edu
Department of Statistics, University of California
Berkeley, CA 94720-4735, USA

Comment 5: Richard A. JOHNSON & Kjell A. DOKSUM

Chen, Lockhart and Stephens (CLS) have developed an asymptotic theory for the linear Box–Cox transformation model where the regression parameters are relatively stable as a function of the transformation parameter. Further, the limit is continuous at the zero value of the vector regression parameter. This is in part an improvement in the asymptotic approach of Bickel & Doksum (1981). By considering asymptotic distributions for certain sequences of parameters, CLS have managed to obtain results that, because of their relative simplicity, shed light on the properties of estimators and tests in the Box–Cox model.

However, the CLS formulation and results have some disconcerting elements. For instance, the transformation parameter is not consistently estimated. Thus, CLS provide more stable scaled regression coefficients, but at the price of knowing less about the transformation scale on which these coefficients are computed. Thus it makes sense to consider an alternative approach to a rigorous asymptotic theory where all of the parameters are fixed but unknown. Such a theory has been provided by Cho, Yeo, Johnson & Loh (2001a). While the Box–Cox transformation cannot produce exactly normal errors, except possibly for $\lambda = 0$, one can maximize the likelihood obtained by assuming that the transformed response $Y^{(\lambda)}$ satisfies the linear model

$$Y^{(\lambda)} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

with independent and identically distributed mean zero normal errors. The large sample theory allows for a nonnormal but mean zero distribution for the errors. Conditions are placed on the densities and consistency is established by first obtaining a uniform strong law of large numbers.

CLS only address parameter estimation and testing, so the search for stable parameters may have some merit. However, if forecasting is the primary purpose of the regression, the quality of the forecast is the proper criterion. Cho, Yeo, Johnson & Loh (2001b) show that asymptotic coverage of the forecast interval attains the nominal level of significance in the fixed parameter setting. A small simulation is included. How does the CLS approach do with forecasting?

One of the reasons CLS consider the framework where $\delta \to 0$ is that for this case (see Section 1.1), where the probability tends to zero that the right-hand side of the Box–Cox model equation takes values that are not possible on the left-hand side. However, an alternative approach would be to extend the Box–Cox transformation so that its domain and range are the whole real line. Such an extension, which preserves convexity ($\lambda > 1$) and concavity ($\lambda < 1$), is provided by Yeo & Johnson (2000).

CLS argue successfully that $\theta = \beta/\sigma$ is more stable than β as a function of λ. However, we can use arguments similar to Brillinger (1983) and Stoker (1986) to argue that when $p \geq 2$, $\beta/|\beta|$, where $| \cdot |$ the Euclidean norm, is perhaps a more stable parameter. Brillinger (1983) asked how stable the least squares estimate $\hat{\beta}_{LS}$ for the linear model

$$Y_i = \alpha + \beta' X_i + \sigma\varepsilon_i$$

would be for the semi-parametric single index model

$$Y_i = g(\alpha + \beta' X_i) + \varepsilon_i, \tag{1}$$

where (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. as (X, Y), $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, ε_i, $i = 1, \ldots, n$, are i.i.d. as ε, $\mathrm{E}\,(\varepsilon) = 0$, $\mathrm{E}\,(\varepsilon^2) = 1$, ε and X are independent, and g is an unknown function. He found that if X has a multivariate normal distribution, then $\hat{\beta}_{LS}$ is consistent for $c\beta$ for some constant c depending on g. Thus for the model (1) with $X \sim \mathsf{N}(\mu, \Sigma)$, $\hat{\beta}_{LS}/|\hat{\beta}_{LS}|$ converges in probability to $\beta/|\beta|$, which does not depend on g. He also found that in this framework,

$$\sqrt{n}\,(\hat{\beta}_{LS} - c\beta) \to \mathsf{N}(0, \Sigma_{LS}),$$

with

$$\Sigma_{LS} = \sigma^2\Sigma^{-1} + \Sigma^{-1}\mathrm{E}\left\{d_0^2(X)(X-\mu)(X-\mu)'\right\}\Sigma^{-1}$$

where $d_0(X) = d(X) - \mathrm{E}\{d(X)\}$ and $d(X) = g(\alpha + \beta' X) - c\beta' X$. Note that the first term in Σ_{LS} is the usual linear model covariance matrix and the second term represents extra variability due to the extra unknown parameter g. Brillinger developed extensions of his results to non-Gaussian as well as fixed (nonrandom) X.

Stoker (1986) extended Brillinger's results to the more general index model where

$$\mathrm{E}\,(Y \mid X) = g(\alpha + \beta' X) \tag{2}$$

for some unknown function g. This model includes the transformation model

$$h(Y) = \alpha + \beta' X + \varepsilon, \quad h \text{ unknown, increasing}$$

where X and ε are independent. Moreover, Stoker introduced the average derivative approach which is based on the parameter vector $\delta = \mathrm{E}\,\{m'(X)\}$, where $m(X) = \mathrm{E}\,(Y \mid X)$ and $m'(X)$ denotes the vector of partial derivatives $\partial m(X)/\partial X_j$, $j = 1, \ldots, p$. Note that whether or not X is Gaussian in model (2), $\delta = c\beta$ for $c = \mathrm{E}\,\{g'(\alpha + \beta' X)\}$, and $\delta/|\delta| = \beta/|\beta|$. Thus $\delta/|\delta|$ is a stable parameter in model (2).

By using integration by parts and the iterated expectation theorem, we find immediately that under regularity conditions (see Stoker 1986),

$$\delta = \mathrm{E}\,\{m'(X)\} = \int m'(x) f(x)\,dx = -\int m(x) f'(x)\,dx = \mathrm{cov}\{Y, \ell(X)\}, \tag{3}$$

where $\ell(X) = -f'(X)/f(X)$ is the location score vector of $f(X)$. Note that both g and m have "disappeared" in the expression $\delta = \mathrm{cov}\{Y, \ell(X)\}$ and that when $X \sim \mathsf{N}(\mu, \Sigma)$, then $\ell(X) = \Sigma^{-1}(X - \mu)$ and

$$\delta = \mathrm{E}\,\{Y\ell(X)\} = \Sigma^{-1}\mathrm{E}\,\{Y(X-\mu)\} = \Sigma^{-1}\Sigma_{XY} = \lim \hat\beta_{LS}, \tag{4}$$

where the limit is in probability. Thus Brillinger's result, which he connected to Stein's identity, follows immediately from (3) and holds in the general index model (2). Note that if we introduce the vector $s(Y) = \mathrm{E}\,(X \mid Y)$, then by (4) and the iterated expectation theorem,

$$X \sim \mathsf{N}(\mu, \Sigma) \;\Rightarrow\; \delta = \mathrm{cov}\{Y, s(Y)\}$$

and we have connected δ to inverse sliced regression (Li 1991).

When X has an unknown distribution, nonparametric estimates $\hat m'(\,\cdot\,)$ and $\hat\ell(\,\cdot\,)$ yield the estimates $\hat\delta = n^{-1}\Sigma w_i \hat m'(X_i)$ and $\tilde\delta = n^{-1}\Sigma w_i Y_i \hat\ell(X_i)$, where the w_i are weights that downweight extreme X_i. Härdle & Stoker (1989), among others, presented conditions that imply the asymptotic normality of $\sqrt{n}\,(\hat\delta - \delta)$ and $\sqrt{n}\,(\tilde\delta - \delta)$. Chaudhuri, Doksum & Samarov (1997) considered estimates obtained by replacing $m(X)$ with the conditional quantile function $q_p(X) = F^{-1}_{Y\,|\,X}(p)$, $0 < p < 1$, and they established $\sqrt{n}$-asymptotic normality of these estimates that also estimate the direction of β in the model (2).

We next look more closely at the implications of the average derivative approach for the transformation model and the Box–Cox estimate. Thus suppose the data are generated by the model

$$h_0(Y) = \alpha_0 + X\beta_0 + \sigma_0\varepsilon, \quad \varepsilon \sim \mathsf{N}(0,1), \tag{5}$$

where $h_0 = h^{(\lambda_0)}$ denotes the true transformation. Here X is random and independent of ε. Similar results can be obtained for the case where X is fixed. Suppose that we work with the wrong, or unknown, transformation $h^{(\lambda)}$ and with the wrong, or unknown, model

$$h^{(\lambda)}(Y) = \alpha(\lambda) + X\beta(\lambda) + \sigma(\lambda)e, \tag{6}$$

where $(\alpha(\lambda), \beta(\lambda), \sigma(\lambda))$ is the probability limit as $n \to \infty$ of the maximum likelihood estimate $(\hat{\alpha}(\lambda), \hat{\beta}(\lambda), \hat{\sigma}(\lambda))$ for the wrong model (6) with $e \sim \mathsf{N}(0,1)$ and

$$e = \left[h^{(\lambda)}(Y) - \left\{\alpha(\lambda) + X\beta(\lambda)\right\}\right]/\sigma(\lambda).$$

Let $g_\lambda(y) = h^{(\lambda)}\{h_0^{-1}(y)\}$; then

$$h^{(\lambda)}(Y) = g_\lambda(\alpha_0 + X\beta_0 + \sigma_0\varepsilon). \tag{7}$$

By differentiating (7), under regularity conditions, we have

$$\mathrm{E}_0\left(\frac{\partial}{\partial X_j} h^{(\lambda)}(Y)\right) = c_\lambda \beta_{0j},$$

where E_0 denotes expected value for the model (5) and

$$c_\lambda = \mathrm{E}_0\left\{g'_\lambda(\mu_0 + X\beta_0 + \sigma_0\varepsilon)\right\}$$

does not depend on the index j. Thus, if we set

$$\delta_j(\lambda) = \mathrm{E}_0\left(\frac{\partial}{\partial X_j} h^{(\lambda)}(Y)\right),$$

then

$$\frac{\delta(\lambda)}{|\delta(\lambda)|} = \frac{\beta_0}{|\beta_0|} \tag{8}$$

and $\delta(\lambda)/|\delta(\lambda)|$ does not depend on λ. This shows that when the true transformation is unknown, the direction of the true regression parameter is identifiable and can be estimated by estimating $\delta(\lambda)/|\delta(\lambda)|$. Moreover, (4) shows that

$$X \sim \mathsf{N}(\mu, \Sigma) \;\Rightarrow\; \frac{\hat{\beta}(\lambda)}{|\hat{\beta}(\lambda)|} \xrightarrow{P} \frac{\beta_0}{|\beta_0|}$$

and, in the framework of Cho, Yeo, Johnson & Loh (2001a), the Box–Cox estimate $\hat{\beta}(\hat{\lambda})$ satisfies $\hat{\beta}(\hat{\lambda})/|\hat{\beta}(\hat{\lambda})| \xrightarrow{P} \beta_0/|\beta_0|$.

Using (6), we find

$$\delta_j(\lambda) = \beta_j(\lambda) + \sigma(\lambda)\mathrm{E}_0\left(\frac{\partial}{\partial X_j} e\right).$$

This, together with (8), shows that when $\mathrm{E}_0(\partial e/\partial X_j)$ is relatively small, e.g., when λ is close to λ_0, then $\beta(\lambda)/|\beta(\lambda)|$ is close to $\delta(\lambda)/|\delta(\lambda)|$ and thus is a stable parameter.

Chen, Lockhart & Stephens actually consider the stability of the "estimate" $\hat{\theta}(\lambda) = \hat{\beta}(\lambda)/\hat{\sigma}(\lambda)$ rather than $\theta(\lambda) = \beta(\lambda)/\sigma(\lambda)$. We can only claim that $\hat{\delta}(\lambda)/|\hat{\delta}(\lambda)|$ is a more stable "estimate" than $\hat{\theta}(\lambda)$ for "n large enough". A Monte Carlo study is needed to compare $\hat{\delta}(\lambda)/|\hat{\delta}(\lambda)|$, $\hat{\beta}(\lambda)/|\hat{\beta}(\lambda)|$, and $\hat{\theta}(\lambda)$ for different n. Here we compare them for Example 2 of Chen, Lockhart & Stephens.

Example 1. We compare the direction vector $\hat{d}(\lambda) = \hat{\beta}(\lambda)/|\hat{\beta}(\lambda)|$ and the CLS signal to noise ratio vector $\hat{\theta}(\lambda) = \hat{\beta}(\lambda)/\hat{\sigma}(\lambda)$ for the Box–Cox textile example (Example 2 in CLS). The result, given in Figure 1, shows that for this example, $\hat{d}(\lambda)$ is much more stable than $\hat{\theta}(\lambda)$. In order to determine a reasonable range for λ in this example, we generated 500 Monte Carlo samples of size $n = 27$ from the model

$$h^{(\lambda_0)}(Y) = \alpha_0 + \sum_{j=1}^{3} \beta_{0j} x_j + \sigma_0\varepsilon, \tag{9}$$

where x_j are as in Box & Cox (1964) and $(\lambda_0, \alpha_0, \beta_{01}, \beta_{02}, \beta_{03}, \sigma_0)$ are the fitted values from Box–Cox (1964). We generated 500 vectors $(\varepsilon_1, \ldots, \varepsilon_{27})$ of i.i.d. $\mathsf{N}(0,1)$ variables and computed 500 vectors $(Y_1, \ldots, Y_{27})$ using the inverse version of (9). For each of the resulting 500 data sets, we computed the Box–Cox $\hat{\lambda}$ and found that these $\hat{\lambda}$ have a nearly symmetric empirical distribution with range from -0.3 to 0.2 and 2.5 and 97.5 percentiles -0.19 and 0.08. These percentiles nearly agree with the Box–Cox posterior distribution 2.5 and 97.5 percentiles -0.20 and 0.08. We found that $\hat{\theta}_j(\hat{\lambda})$, $j = 1, 2, 3$, have slightly skew distributions while $\hat{d}_j(\lambda)$, $j = 1, 2, 3$, have nearly symmetric distributions. Table 5-1 gives the Monte Carlo standard errors (SE) of $\hat{\theta}_j(\hat{\lambda})$ and $\hat{d}_j(\hat{\lambda})$, $j = 1, 2, 3$, together with their Monte Carlo skewness coefficients (SC). Table 5-2 gives the λ "known" case.

TABLE 5-1: Standard errors (SE) and skewness coefficients (SC) of $\hat{\theta}_j(\hat{\lambda})$ and $\hat{d}_j(\hat{\lambda})$, $j = 1, 2, 3$.

	$\hat{\theta}_1(\hat{\lambda})$	$\theta_2(\hat{\lambda})$	$\hat{\theta}_3(\lambda)$	$\hat{d}_1(\hat{\lambda})$	$\hat{d}_2(\hat{\lambda})$	$\hat{d}_3(\hat{\lambda})$
Monte Carlo SE	0.880	0.705	0.499	0.024	0.031	0.035
CLS Approximate SE	0.664	0.526	0.376			
SC	0.668	-0.522	-0.691	-0.084	0.114	0.111

TABLE 5-2: Monte Carlo standard errors (SE) and skewness coefficients (SC) of $\hat{\theta}_j(\lambda)$ and $\hat{d}_j(\lambda)$ with $\lambda = -0.06, j = 1, 2, 3$.

	$\hat{\theta}_1(\lambda)$	$\hat{\theta}_2(\lambda)$	$\hat{\theta}_3(\lambda)$	$\hat{d}_1(\lambda)$	$\hat{d}_2(\lambda)$	$\hat{d}_3(\lambda)$
SE	0.878	0.700	0.449	0.023	0.029	0.034
SC	0.557	-0.647	-0.359	-0.252	-0.033	0.086
CLS Approx.	0.664	0.526	0.376			

These tables show that the CLS approximations are fairly good even though the sample size is only $n = 27$. However $\hat{d}_j(\,\cdot\,)$ has the advantage of having a nearly symmetric distribution in addition to being more stable (Figure 5-1).

Next, we point out that by using the rank likelihood, it is possible to construct a 100% stable estimate of $\theta = \beta/\sigma$ when the data are generated by the model

$$h(Y_i) = \alpha + \beta' x_i + \sigma\varepsilon_i, \quad h \text{ unknown, increasing} \tag{10}$$

where now the vectors x_i are non-random, $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. and ε_i has a density f_ε which does not depend on x_i. Let $R_i = R(Y_i)$ denote the rank of Y_i, then

$$R_i = R(Y_i) = R\{h(Y_i)\} = R(\theta' x_i + \varepsilon_i).$$

By Hoeffding's formula, the rank (partial) likelihood based on $R = (R_1, \ldots, R_n)'$ is

$$L_R(\theta) = \mathrm{P}\,(R = r) = \mathrm{E}\left\{\prod_{i=1}^{n} \frac{f_\varepsilon(V^{(r_i)} - \theta' x_i)}{f_\varepsilon(V^{(r_i)})}\right\},$$

where $V^{(1)} \leq \ldots \leq V^{(n)}$ are the order statistics of a sample from f_ε. Now $\hat{\theta}_R = \text{argmax } L_R(\theta)$ does not depend on h and is 100% stable. Here f_ε is specified, and when $\varepsilon \sim \mathsf{N}(0,1)$, then (10) generalizes the Box–Cox model to arbitrary increasing transformations.

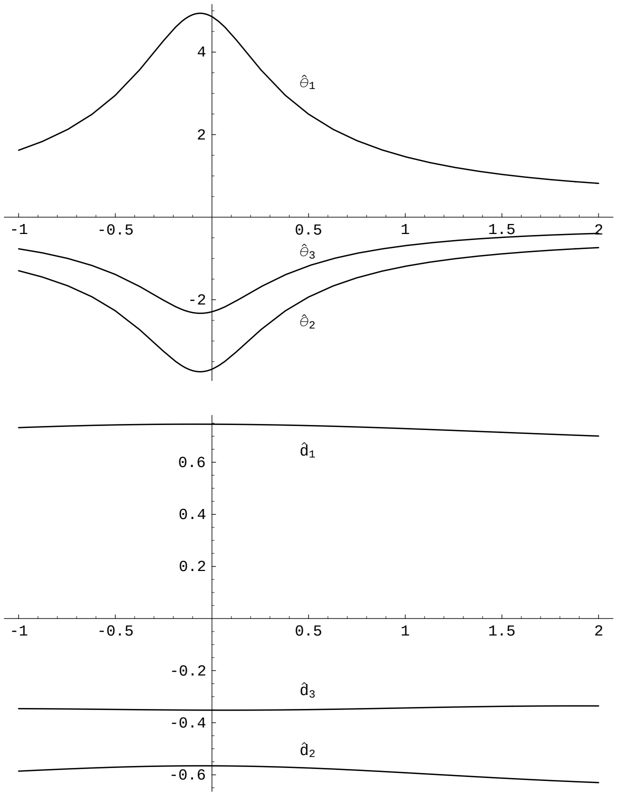

FIGURE 5-1: Stability of $\hat{\theta}(\lambda)$ and $\hat{d}(\lambda)$.

For this model, Doksum (1987) used Monte Carlo methods to compute $L_R(\theta)$ and $\hat{\theta}_R$, and he considered properties of estimates based on ranks for a model with

$$\theta \in \Omega = \left\{ \theta : \sum_{i=1}^{n} (\theta' x_i)^2 \leq K, \ \max|\theta' x_i| \to 0 \right\},$$

where the x_i are centered. Note that Ω is a subset of the parameter sequences of CLS. Bickel & Ritov (1997) established a locally asymptotically normal structure for $L_R(\theta)$ that can be used to construct Monte Carlo methods with better convergence properties than those in Doksum (1987).

ACKNOWLEDGEMENT

The authors are grateful to Richard Bourgon for producing Figure 5-1 and Tables 5-1 and 5-2.

Richard A. JOHNSON: rich@stat.wisc.edu
Department of Statistics, University of Wisconsin
1210 West Dayton Street, Madison, WI 53706-1613, USA

Kjell A. DOKSUM: doksum@stat.berkeley.edu
Department of Statistics, University of California
Berkeley, CA 94720-4735, USA

Comment 6: Peter M. HOOPER

Inferential problems associated with the Box–Cox transformation have generated a variety of proposed solutions in the statistical literature. The issues are both philosophical and technical in nature. For example: What inferential questions are of scientific interest following estimation of the transformation parameter λ? What parameters are best suited for addressing such questions? Are standard tests and confidence intervals still applicable, or are new methods required? How do we account for the fact that the scale of the transformed response varies with λ? And how should an asymptotic analysis deal with the truncation effect inherent in power transformations?

Professors Chen, Lockhart, and Stephens employ two key ideas in addressing these issues. First, they consider inferential questions associated with a parameter vector $\theta = \beta/\sigma$ defined so that its scale remains stable as λ varies. Second, they employ an asymptotic framework where a "truncation effect" parameter δ approaches zero as the sample size n becomes large. Most of my comments concern the appropriateness of θ as an estimand, but I will begin with a few words about the asymptotics.

The authors argue that "a good sequence of models should have computable limiting distributions for the estimates and should have finite sample distributions which converge quickly to their limits". These are appropriate goals, and the approach adopted by the authors seems to achieve these goals. I expect that, in most applications, the truncation effect is small enough to be ignored. An asymptotic framework where $\delta \to 0$ as $n \to \infty$ provides a useful formulation of this underlying expectation. The scope of the asymptotic results is impressive. The variance formulae provide insight about how the estimates $\hat\lambda$ and $\hat\theta$ are affected by the linear model structure (e.g., regression or single-factor ANOVA) and by the error (large or small σ, departures from normality). The tests of normality are another important contribution.

The asymptotic theory is used to construct tests and confidence intervals for linear functions of θ. I have serious doubts about the practical relevance of the confidence intervals. I believe the intervals are technically correct, i.e., coverage probabilities should be close to nominal levels. My concerns are with (i) a variance inflation effect related to the estimation of σ and (ii) the interpretation of θ as an estimand.

The authors' primary motivation for introducing θ is to reduce the variance in $\hat\beta$ related to its correlation with $\hat\lambda$. Dividing β by σ achieves this aim, nearly eliminating this source of variance in most applications. Unfortunately, by estimating the denominator σ, one introduces another source of variance, the term $(2n)^{-1}\theta\theta'$ in the authors' expression (17) for $V(\hat\theta)$. This term becomes unbounded as $\sigma \to 0$ with n and β fixed, a situation where one would expect increasingly narrow confidence intervals for regression coefficients. In typical applications the $(2n)^{-1}\theta\theta'$ term can be large compared with $(X_1'X_1)^{-1}$, e.g, compare SE$_1$ with SE$_2$ in the authors' Tables 4 and 6.

To better understand how the variance of $\hat\theta$ is related to estimation of σ and to see that this has nothing to do with the Box–Cox transformation, I found it helpful to examine the mean square error of $\hat\theta$ in the much simpler context of a normal-theory linear model with no transformation of the response. Suppose we have $Y \sim \mathsf{N}_n(\mu 1 + X_1\beta, \sigma^2 I_n)$, where the p columns of X_1 are centered and the responses are *not* necessarily positive. Let s^2 be the usual unbiased estimator of σ^2. We then have $\hat\beta$ and s^2 independent, $\hat\beta \sim \mathsf{N}_p(\beta, \sigma^2(X_1'X_1)^{-1})$, and $\nu s^2/\sigma^2 \sim \chi^2_\nu$, where $\nu = n - p - 1$. Express $\hat\theta - \theta$ as a sum of two uncorrelated random vectors,

$$\hat\theta - \theta = s^{-1}(\hat\beta - \beta) + (s^{-1} - \sigma^{-1})\beta. \tag{11}$$

Using the moments of the Gamma distribution, one may then calculate

$$\mathrm{E}\left\{(\hat\theta - \theta)(\hat\theta - \theta)'\right\} = \frac{\nu}{\nu - 2}(X_1'X_1)^{-1} + \frac{c(\nu)}{2\nu}\theta\theta', \tag{12}$$

where

$$c(\nu) = \frac{4\nu(\nu-1)}{\nu-2} - \frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}(2\nu)^{3/2}.$$

Furthermore, $c(\nu) \to 1$ as $\nu \to \infty$, and the asymptotic bias of $\hat{\theta}$ is $3\theta/(4\nu)$. For fixed n, σ and X_1, the variance of $a'\hat{\beta}$ remains fixed as $a'\beta$ varies, but the variance of $a'\hat{\theta}$ increases with $(a'\theta)^2$ at a rate proportional to $(a'\theta)^2$.

The interpretation of θ also presents difficulties. The regression coefficient vector β quantifies the rate of change in $\mathrm{E}\,\{Y(\lambda)\}$ associated with changes in the predictors. The interpretation of this rate depends on the scale of both $Y(\lambda)$ and the predictors. Since the scale of $Y(\lambda)$ depends on λ, the scientific relevance of β is unclear because λ is unknown. Working with the scaled coefficient vector θ appears to avoid this problem. The estimates $\hat{\lambda}$ and $\tilde{\sigma}(\hat{\lambda})$ vary together so that $Y(\lambda)/\tilde{\sigma}(\lambda)$ remains stable under variation in λ. Changes in the scaled transformed response can thus be interpreted as if $\lambda = \hat{\lambda}$. Unfortunately, a new problem arises. The scaled estimand describes the rate of change of the transformed response proportional to the error standard deviation σ. Again, the scientific relevance of θ is unclear because σ is unknown. It is difficult to evaluate "interpretation" or "scientific relevance" mentioned in the abstract. The reader may wish to recall real applications of linear models from his/her own experience and consider whether the estimation of θ seems useful in those contexts. I can think of no application from my own experience where I would want to estimate θ.

One may contrast inference for θ with a well-known data-based scaling technique. Box & Cox (1964) suggested that effects be analysed in terms of an approximate linear model for the scaled transformed responses $Y(\lambda)/\dot{y}^{\lambda-1}$, where $\dot{y}$ is the geometric mean. The scaled coefficient estimates $\hat{\beta}_j(\lambda)/\dot{y}^{\lambda-1}$ are usually stable with respect to changes in λ (Duan 1993) and hence can be interpreted as if $\lambda = \hat{\lambda}$. The scale factor $\dot{y}^{\hat{\lambda}-1}$ provides a known reference for interpreting the scaled coefficients.

My own preference when analysing regression data is to regard transformation of the response as model selection and to condition the definition of the regression coefficients on the model selected. This approach is investigated in Hooper & Yang (1997). Following Cohen & Sackrowitz (1987), we applied a least squares criterion to define the function $\beta(\ell) = (X_1'X_1)^{-1}X_1'\mathrm{E}\,\{Y(\ell)\}$ for all possible values ℓ of λ. We thus have $\beta(\lambda) = \beta$ for the true λ. Our estimand $\beta(\hat{\lambda})$ is easily interpreted since the scale of $Y(\hat{\lambda})$ is known. Our estimand is not equivalent to $\mathrm{E}\,(\hat{\beta} \mid \hat{\lambda})$, although our point of view is similar to that of Hinkley & Runger (1984). Variation in $\hat{\beta} - \beta(\hat{\lambda})$ is not subject to the variance inflation affecting $\hat{\beta} - \beta(\lambda)$, and standard confidence intervals (ignoring estimation of λ) are applicable. The situation is simplest when β is large, σ is small, and $\hat{\lambda}$ is determined primarily by the deterministic component of the model (see Section 3.1 in our paper). Here $\hat{\lambda} - \lambda$ behaves much like an estimated regression coefficient for an additional predictor—Atkinson (1985, Section 6.4) describes inferential techniques based on this idea—and the conditional (given $\hat{\lambda}$) coverage probability of a confidence interval for $a'\beta(\hat{\lambda})$ is typically close to the nominal level. By contrast, this is a situation where $a'\hat{\theta}$ is strongly affected by variance inflation.

My critique of θ as an estimand does not apply to its use in tests of null hypotheses. Tests of linear hypotheses of the form $\mathcal{H}_0 : A\theta = 0$ are equivalent to corresponding hypotheses about β, so there are no problems with interpretation. Variance inflation is also not a problem because $A\theta\theta'A'$ vanishes (under the null hypothesis). The asymptotic distribution theory for $\hat{\theta}$ provides a rigorous foundation for tests of linear hypotheses.

Peter M. HOOPER: hooper@math.ualberta.ca
Department of Mathematical Sciences, The University of Alberta
Edmonton, Alberta, Canada T6G 2G1

Comment 7: Zhenlin YANG

1. INTRODUCTION

The effect of estimating a transformation on the subsequent inferences is an important issue in the applications of the Box–Cox transformation method. Many have made their contributions to this issue, directly or indirectly. Those include, among others, Bickel & Doksum (1981), Carroll & Ruppert (1981), Carroll (1982), Doksum & Wong (1983), Hinkley & Runger (1984), Taylor (1989), Duan (1993) and Yang (1999).

The present authors have studied this issue by concentrating on the ratio of the regression slope and the error standard deviation and provide important results to show that the estimation of this scaled slope is much more stable than the estimation of the slope itself with respect to the transformation estimation. Making use of some of the results of Yang (1999), I am able to provide some similar results under a slightly different set-up. Also, I carry out some Monte Carlo experiments to investigate the finite sample effect of transformation estimation. Both reinforce the authors' findings.

2. THE GENERAL RESULTS

First, I introduce some different notation to reflect the changes in the set-up: $\mathbf{Y}(\lambda_0) = \mathbf{X}\beta_0 + \sigma_0\varepsilon$, $\eta_i = x_i'\beta_0$ and $\delta_i = \lambda_0\sigma_0/(1+\lambda_0\eta_i)$. Thus β_0, and hence $\theta = \beta_0/\sigma_0$, includes the intercept parameter and the definition of δ_i incorporates the values of the regressors. Let $\xi_0 = (\beta_0', \lambda_0, \sigma_0)$ and let $\hat{\xi}$ be an M-estimator of ξ_0 that solves the following estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \Psi_i(Y_i; \hat{\xi}) = 0_{(p+3)\times 1},$$

where the function Ψ_i is partitioned according to β_0, λ_0, and σ_0. Define

$$\overline{\Psi} = \frac{1}{n}\sum_{i=1}^{n} \Psi_i(Y_i, \xi_0),$$

and $\mathbf{A} = \mathrm{E}\,(\partial\bar{\Psi}/\partial\xi_0)$, both partitioned accordingly. The elements of $\overline{\Psi}$ are denoted by $\overline{\Psi}_i$, $i = 1,2,3$, and the elements of $\mathbf{A}$ by $\mathbf{A}_{ij}$, $i,j = 1,2,3$. Let $\tilde{\beta}$ and $\tilde{\sigma}$ be, respectively, the M-estimate of β_0 and σ_0 when λ_0 is known. Assume that the conditions C1, C2, C3, and C5 of Yang (1999) are satisfied. Following Taylor expansions on the estimating equation, we find

$$\hat{\beta} - \beta_0 = -\mathbf{A}_{11}^{-1}\overline{\Psi}_1 - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\hat{\lambda} - \lambda_0) + O_p(n^{-1}), \tag{13}$$

$$\hat{\sigma} - \sigma_0 = -\mathbf{A}_{33}^{-1}\overline{\Psi}_3 - \mathbf{A}_{33}^{-1}\mathbf{A}_{32}(\hat{\lambda} - \lambda_0) + O_p(n^{-1}), \tag{14}$$

$$\hat{\lambda} - \lambda_0 = \frac{\overline{\Psi}_2 - \mathbf{A}_{23}\mathbf{A}_{33}^{-1}\overline{\Psi}_3 - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\overline{\Psi}_1}{\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} - \mathbf{A}_{22} + \mathbf{A}_{23}\mathbf{A}_{33}^{-1}\mathbf{A}_{32}} + O_p(n^{-1}). \tag{15}$$

Equating $\hat{\lambda}$ to λ_0 in (1) and (2) gives

$$\tilde{\beta} - \beta_0 = -\mathbf{A}_{11}^{-1}\overline{\Psi}_1 + O_P(n^{-1})$$

and

$$\tilde{\sigma} - \sigma_0 = -\mathbf{A}_{33}^{-1}\overline{\Psi}_3 + O_P(n^{-1}).$$

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

$$\hat{\sigma}^{-1} = \tilde{\sigma}^{-1} - \sigma_0^{-2}(\hat{\sigma} - \tilde{\sigma}) + O_p(n^{-1}) = \tilde{\sigma}^{-1} + \sigma_0^{-2}\mathbf{A}_{33}^{-1}\mathbf{A}_{32}(\hat{\lambda} - \lambda_0) + O_p(n^{-1}). \tag{16}$$

Writing (1) as $\hat\beta = \tilde\beta - \mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\hat\lambda - \lambda_0) + O_p(n^{-1})$ and combining this with (4), we find

$$\hat\theta = \tilde\theta + (\theta\mathbf{A}_{33}^{-1}\mathbf{A}_{32} - \mathbf{A}_{11}^{-1}\mathbf{A}_{12})(\hat\lambda - \lambda_0)\sigma_0^{-1} + O_p(n^{-1}). \tag{17}$$

The second term in the expansion (5) reflects the effect of the estimating transformation. The magnitude of this effect can be studied in detail if the estimating function Ψ_i and the transformation function $Y(\lambda)$ are both specified. This result can be compared with the expansion for $\hat\beta$ to see which quantity is more stable with respect to $\hat\lambda$.

3. THE BOX–COX POWER TRANSFORMATION AND NEAR NORMAL ERRORS

When the Box–Cox power transformation is used and the errors are approximately normal, the estimating function corresponding to the maximum likelihood estimation takes the form

$$\Psi_i(Y_i,\xi_0) = \left\{\begin{array}{lcl} \Psi_{1i}(Y_i,\xi_0) & = & \sigma_0^{-2}x_i\{Y_i(\lambda_0) - x_i'\beta_0\}, \\ \Psi_{2i}(Y_i,\xi_0) & = & \log Y_i - \sigma_0^{-2}\{Y_i(\lambda_0) - x_i'\beta_0\}\dot Y_i(\lambda_0), \\ \Psi_{3i}(Y_i,\xi_0) & = & \sigma_0^{-3}\{Y_i(\lambda_0) - x_i'\beta_0\}^2 - \sigma_0^{-1}, \end{array}\right.$$

where

$$\dot Y_i(\lambda_0) = \partial Y_i(\lambda_0)/\partial\lambda_0 = \left\{\begin{array}{ll} \lambda_0^{-1}\{1 + \lambda_0 Y_i(\lambda_0)\}\log Y_i - \lambda_0^{-1}Y_i(\lambda_0), & \lambda_0 \neq 0, \\ \frac{1}{2}(\log Y_i)^2, & \lambda_0 = 0. \end{array}\right.$$

With the Box–Cox power transformation, the exact normality of $Y_i(\lambda_0)$ is incompatible with the positivity of Y_i. Hence there is a truncation effect when the above Ψ_i function is used to approximate to the true likelihood estimating function. However, this truncation effect is small if $\delta_0 = \max|\delta_i|$ is small, which is achievable when (i) σ_0 is small, (ii) λ_0 is small, or (iii) $\min|\eta_i|$ is large.

Assuming that the truncation effect is small and using the Ψ_i function given above, one can easily evaluate all the quantities involved in (1)–(3) to simplify the expansions. Further, when $\lambda_0 = 0$, all the expansions can be expressed explicitly in terms of ξ_0 and $\varepsilon_i s$, which allows one to examine the affecting term in detail as well as to find an explicit expression for the variance inflation. When $\lambda_0 \neq 0$, however, an approximation to $\log Y_i$ is necessary. When δ_0 is small, we have,

$$\lambda_0\log Y_i = \log(1 + \lambda_0\eta_i) + \delta_i\varepsilon_i - \tfrac{1}{2}\delta_i^2\varepsilon_i^2 + O_p(\delta_i^3). \tag{18}$$

I use ε^2 to mean $(\varepsilon_i^2)_{n\times1}$, etc. Thus,

$$\dot{\mathbf{Y}}(\lambda_0) = \left\{\begin{array}{ll} \frac{1}{2}(\eta^2 + 2\eta\sigma_0\varepsilon + \sigma_0^2\varepsilon^2), & \lambda_0 = 0, \\ \dfrac{1}{\lambda_0^2}[(\mathbf{1} + \lambda_0\eta)\#\phi] + \dfrac{\sigma_0}{\lambda_0}\phi\#\varepsilon + \dfrac{\sigma_0}{2\lambda_0}\delta\#\varepsilon^2 - \dfrac{\sigma_0}{2\lambda_0}\delta^2\#\varepsilon^3 + O_p(\delta^3), & \lambda_0 \neq 0. \end{array}\right.$$

By assuming the first six moments of ε_1 are the same as those of a standard normal random variable and making use of the above approximations, Yang (1999) derived the following explicit expansions:

$$\begin{array}{rcl} \dfrac{\hat\lambda - \lambda_0}{\sigma_0} & = & \dfrac{-\frac{1}{2}(\mathbf{M}\eta^2)'\varepsilon - \sigma_0(\eta - \bar\eta\mathbf{1})'\varepsilon^2 + \frac{1}{2}\sigma_0^2\mathbf{1}'(3\varepsilon - \varepsilon^3)}{\frac{1}{4}||\mathbf{M}\eta^2||^2 + 2\sigma_0^2||\eta - \bar\eta\mathbf{1}||^2 + \frac{3}{2}n\sigma_0^4} + O_p(n^{-1}), \quad \lambda_0 = 0, \\ \dfrac{\hat\lambda - \lambda_0}{\lambda_0} & = & \dfrac{-(\delta^{-1}\#\phi + \frac{1}{2}\delta)'\mathbf{M}\varepsilon - (\phi - \bar\phi)'\varepsilon^2 + \frac{1}{2}(3\varepsilon - \varepsilon^3)}{||\mathbf{M}(\delta^{-1}\#\phi + \frac{1}{2}\delta)||^2 + 2||\phi - \bar\phi\mathbf{1}||^2 + \frac{3}{2}||\delta||^2} + O_p(n^{-1}) + O_p(\delta_0^3),\ \lambda_0 \neq 0, \end{array}$$

where $\phi = \log(1 + \lambda_0\eta)$ and $||\cdot||$ denotes the Euclidean norm. These lead immediately to the approximations to the variance of $\hat\lambda$ for large n and small δ_0,

$$\begin{array}{rcl} \tau^2(\xi_0) \equiv \mathrm{var}(\hat\lambda) & \approx & \sigma_0^2\{\frac{1}{4}||\mathbf{M}\eta^2||^2 + 2\sigma_0^2||\eta - \bar\eta\mathbf{1}||^2 + \frac{3}{2}n\sigma_0^4\}^{-1}, \quad \lambda_0 = 0, \\ \tau^2(\xi_0) \equiv \mathrm{var}(\hat\lambda) & \approx & \lambda_0^2\{||\mathbf{M}(\delta^{-1}\#\phi + \frac{1}{2}\delta)||^2 + 2||\phi - \bar\phi\mathbf{1}||^2 + \frac{3}{2}||\delta||^2\}^{-1}, \quad \lambda_0 \neq 0. \end{array}$$

Now,

$$\mathbf{A}_{11} = -\frac{1}{n\sigma_0^2}\mathbf{X}'\mathbf{X}, \quad \mathbf{A}_{12} = \frac{1}{n\sigma_0^2}\mathbf{X}'\mathrm{E}\{\dot{\mathbf{Y}}(\lambda_0)\}, \quad \mathbf{A}_{33} = \frac{-2}{\sigma_0^2}, \quad \mathbf{A}_{32} = \frac{2}{n\sigma_0^2}\,\mathrm{E}\{\varepsilon'\dot{\mathbf{Y}}(\lambda_0)\}.$$

With the approximation (6) and the explicit expansion for $\hat{\lambda}$, the expansion (5) becomes

$$\hat{\theta} = \tilde{\theta} + v(\xi_0)(\hat{\lambda} - \lambda_0)\sigma_0^{-1} + O_p(n^{-1}) + O_p(\theta_0^3), \tag{19}$$

where

$$v(\xi_0) = \begin{cases} \frac{1}{2}\mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\big\{(\eta - \bar{\eta}\mathbf{1})^2 - (\bar{\eta}^2 - \sigma_0^2)\mathbf{1}\big\}, & \lambda_0 = 0, \\ \lambda_0^{-1}\mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\big\{\lambda_0^{-1}(\mathbf{1} + \lambda_0\eta)\#\phi - \big(1 + \bar{\phi} - \frac{3}{2}\bar{\delta^2}\big)\eta + \frac{1}{2}\sigma_0\delta\big\}, & \lambda_0 \neq 0. \end{cases}$$

Finally, considering $\tilde{\sigma}^{-1}$ as a function of $\tilde{\sigma}^2$ and Taylor expanding it around σ_0^2 give

$$\tilde{\sigma}^{-1} = \sigma_0^{-1} + \frac{1}{2\sigma_0^3}(\tilde{\sigma}^2 - \sigma_0^2) + O_p(n^{-1}) = \sigma_0^{-1} + \frac{1}{n\sigma_0}\sum_{i=1}^{n}(\varepsilon_i^2 - 1) + O_p(n^{-1}).$$

This, combined with

$$\tilde{\beta} = \beta_0 - \mathbf{A}_{11}^{-1}\mathbf{\Psi}_1 + O_p(n^{-1}) = \beta_0 + \sigma_0\mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\varepsilon + O_p(n^{-1})$$

gives

$$\tilde{\theta} = \theta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon + \frac{\theta}{n}\sum_{i=1}^{n}(\varepsilon_i^2 - 1) + O_p(n^{-1})$$

These lead to approximations for the variances of $\hat{\theta}$ and $\tilde{\theta}$ when n is large and δ_0 is small

$$\mathbf{V}(\hat{\theta}) \approx (\mathbf{X}'\mathbf{X})^{-1} + \frac{1}{2n}\theta\theta' + \frac{1}{\sigma_0^2}v(\xi_0)v(\xi_0)'\tau^2(\xi_0)$$

and $\mathbf{V}(\tilde{\theta}) \approx (\mathbf{X}'\mathbf{X})^{-1} + \frac{1}{2n}\theta\theta'$. Similarly, one obtains the expansions for $\hat{\beta}$ and its variance, namely

$$\hat{\beta} = \tilde{\beta} + w(\xi_0)(\hat{\lambda} - \lambda_0) + O_p(n^{-1})$$

and

$$\mathbf{V}(\hat{\beta}) \approx \sigma_0^2(\mathbf{X}'\mathbf{X})^{-1} + w(\xi_0)w(\xi_0)'\tau^2(\xi_0),$$

where

$$w(\xi_0) = \begin{cases} \frac{1}{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\eta^2 + \sigma_0^2\mathbf{1}), & \lambda_0 = 0, \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big\{\lambda_0^{-2}(\mathbf{1} + \lambda_0\eta)\#\phi - \lambda_0^{-1}\eta + \frac{1}{2}\sigma_0\lambda_0^{-1}\delta\big\}, & \lambda_0 \neq 0. \end{cases}$$

The expansions for the $\lambda_0 = 0$ case are the same as those given by the authors, except that the intercept parameter is also included here. The effect of estimating the transformation on the estimation of θ is governed by $v(\xi_0)$, and that on the estimation of β_0 is governed by $w(\xi_0)$. A close examination of the two quantities for the case of $\lambda_0 = 0$ reveals that there are many cases in which the component of $v(\xi_0)$ corresponding to the slope parameters vanishes (see the detailed discussions provided by the authors), but no such cases exist for $w(\xi)$. In the case of $\lambda_0 \neq 0$, the two quantities behave similarly to the case of $\lambda_0 = 0$. This means that the effect on the estimation of θ can be zero or small, but the effect on the estimation of β_0 is generally large. Note that the effect of the estimating transformation on the estimation of the (scaled) intercept parameter is not small in general. When the regressors are centered, $\bar{\eta} = \mu_0$.

4. MONTE CARLO SIMULATION

I now present some Monte Carlo simulation results to show the finite sample effect of the estimating transformation. The model used in the simulation is: $Y_i(\lambda_0) = \mu_0 + \beta_0 x_i + \sigma_0 \varepsilon_i$, $i = 1, \ldots, n$. The x values are centered values of $\exp(2i/n)$, $i = 1, \ldots, n$. I use the relative bias (RB) and the relative efficiency (REF) to measure the effects. The simulation results are summarized in Table 7-1. From the results, we see that there is generally a large effect on the estimation of β_0, but a very small effect on the estimation of β_0/σ_0. It is interesting to note that as the sample size increases, the effect on the estimation of β_0/σ_0 reduces significantly, but the effect on the estimation of β_0 does not seem to change much. Estimating the transformation induces extra bias in the estimation of β_0/σ_0, but only slightly. Additional simulations (not reported here) reveal that the spread in the x values matters. A larger spread gives a smaller effect on the estimation of β_0/σ_0. The magnitude of β_0/σ_0 affects the value of MSE, but affects very little on the relative efficiency.

TABLE 7-1: A summary of simulated RB and REF, where, for example, RB($\hat{\theta}$) $= 100 \times (\hat{\theta} - \theta)/\theta$ and REF($\tilde{\theta}, \hat{\theta}$) $= \mathrm{MSE}(\hat{\theta})/\mathrm{MSE}(\tilde{\theta})$. The MSE stands for the mean squared error.

			Relative bias				Relative efficiency		
ℓ	σ_0	RB($\tilde{\beta}$)	RB($\hat{\beta}$)	RB($\tilde{\theta}$)	RB($\hat{\theta}$)	MSE($\tilde{\beta}$)	REF($\tilde{\beta}, \hat{\beta}$)	MSE($\tilde{\theta}$)	REF($\tilde{\theta}, \hat{\theta}$)
		$n = 20$,	$\mu_0 = 8.0$,	$\beta_0 = 2.0$					
.25	.05	0.00	0.02	10.18	13.77	.000036	81.14	79.6946	1.28
	.1	0.00	0.08	9.63	13.18	.000143	80.52	18.5465	1.32
	.5	0.01	4.49	10.11	13.76	.003639	80.07	0.7933	1.30
.0	.05	0.00	0.02	10.40	13.93	.000035	30.23	80.3997	1.29
	.1	0.00	0.09	10.09	13.60	.000143	30.26	19.4002	1.27
	.5	0.01	2.57	10.08	13.65	.003565	29.34	0.7782	1.28
		$n = 50$,	$\mu_0 = 8.0$,	$\beta_0 = 2.0$					
.25	.05	0.00	0.03	3.64	4.76	.000015	82.74	20.9905	1.11
	.1	0.00	0.14	3.73	4.87	.000061	83.14	5.3149	1.11
	.5	−0.03	1.95	3.53	4.72	.001512	76.03	0.2155	1.13
.0	.05	0.00	0.02	3.68	4.80	.000015	31.16	20.8424	1.11
	.1	−0.01	0.04	3.66	4.83	.000060	32.13	5.2787	1.12
	.5	0.02	1.05	3.71	4.92	.001494	28.64	0.2202	1.13
		$n = 100$,	$\mu_0 = 8.0$,	$\beta_0 = 2.0$					
.25	.05	0.00	0.00	1.76	2.29	.000008	85.55	9.1725	1.05
	.1	0.00	0.04	1.87	2.41	.000030	87.02	2.3389	1.06
	.5	0.02	0.84	1.87	2.42	.000768	73.57	0.0951	1.08
.0	.05	0.00	0.01	1.84	2.37	.000008	31.15	9.3085	1.06
	.1	0.00	0.01	1.73	2.28	.000031	32.75	2.2728	1.06
	.5	−0.02	0.59	1.73	2.30	.000769	27.81	0.0942	1.08

Zhenlin YANG: ecsyzl@nus.edu.sg
Department of Economics, National University of Singapore
1 Arts Link, Singapore 117570

Rejoinder: Gemai CHEN, Richard A. LOCKHART & Michael A. STEPHENS

We should first like to thank “the two Christians”, Professors Léger and Genest, for the excellent organisation of the read session of this paper in Montréal; and then we thank the discussants for the care they have given to their remarks, and for the very extensive range of their comments.

Several of the discussants (Professors Reid, Johnson and Doksum, Hooper and Yang) have in effect suggested alternative formulations of the analysis, and Professors Doksum and Johnson, and Professor Yang, have given considerable detail. We shall return to these below, but first we concentrate on replying to specific concerns and criticisms raised by the discussants.

1. *Value of δ.* The first concern is the role played by δ, discussed by Professors Abadir, Bickel and Hooper. For our purposes, the size of δ is a guide to when we can apply our analysis to model (2) of the paper. A small δ implies that the chances of finding a value given by the right-hand side (for any likely values of x and ε) which cannot give a $Y(\lambda)$ are remote. Suppose, for simplicity, that all the values of x are actually at their respective means; then an “impossible” (i.e., negative) $Y(\lambda)$ requires an $\varepsilon < -1/|\delta|$. For our Example 1, $\delta = 0.097$ and the chance is less than $\Phi(-10.3) = 0$ to 20 decimal places; for the Box–Cox textile example, $1/|\delta| = 90.9$, so that the possibility of negative Y is again negligible; on the other hand, if no transformation is made, $1/|\delta| = 1.75$, giving a probability of 0.04 of incompatibility.

Thus, although some discussants have proposed methods of dealing with the possibility of a negative Y, these considerations are unnecessary in practice, for our analysis with our data sets. Nevertheless, there will be some data sets where one must decide how small is small, and Professor Abadir raises the question of the impact on hypothesis testing for other parameters if this step is viewed as a preliminary test. It would be useful to investigate this point further. We thank Professor Abadir also for the references to the economics literature.

2. *Choice of λ and subsequent inference.* The question of whether or not to allow for estimation of λ in subsequent inference has been central to the controversy surrounding the Box–Cox transformation. We stand chastened by Professor McCullagh’s observation that the Box–Cox paper “does not contain a recommendation, or even a suggestion that $\hat{\lambda}$ be treated as the true value for the purpose of further analysis”; this demonstrates that over the 12 years this paper has been in the works, we did not check enough on the original source. But the check has now been made, and it is clear that Professors Box and Cox do expect to “fix one, or possibly a small number, of λs and go ahead with the detailed estimation . . .” (Box & Cox 1964, p. 213, line 15 ff.)—but the choice will not necessarily be the MLE of λ, as we wrote. In fact, Professors Box and Cox, in both of their examples, use what we have called “snap to the grid” (STG) to decide on the λ in the transformation; that is, to take a “sensible” λ from a possible list of values, basing the decision on scientific grounds, or, as in the Box–Cox toxicity example, on grounds of interpretability.

In this connection, we recall the story of Cavendish, who was investigating the law of attraction between magnetic poles. (We shall update the story to fit the present context). Thus we suppose his graduate student did the work and reported that the MLE of the inverse power of the distance between the poles was 1.98. The distinguished scientist replied that God did not deal with powers like 1.98, and chose the number 2 instead. The rest, as they say, is history. (This story has also been credited to Gauss. In Stephens (1975), “God” was changed to “Nature”, as a referee objected that the original might offend some people’s religious susceptibilities. So far, Naturalists have not been offended.)

STG as model selection. Suppose that STG is used in the Box–Cox–Cavendish manner, with the data being regarded as simply “pointing the way” to the (true?) λ; then we feel there is some force to Professor Hooper’s point of view that the choice is simply model selection. Professor Reid also regards the STG technique as "most sensible", as do we. Then the implication appears to be that subsequent inference is made as though λ were fixed, following Professors Box and Cox (see their examples). Similarly, tests for normality would take the usual form using the transformed residuals, as suggested by Professor Reid; that is, the final term in our $\rho(s, t)$ (Section 5.1) will be omitted, and the test proceeds as described in Stephens (1986, Section 4.8.5).

The Box–Cox transformation more generally. Unfortunately, as we know, it is often the case in deciding on a model (not necessarily in the Box–Cox context) that several models or distributions are tried before one is adopted for fitting data, and the conditionality on the final choice is then paramount, although often not admitted and certainly not accounted for. (For example, a distribution may be preferred and decided upon after calculating several goodness-of-fit statistics and choosing one which does not reject). We therefore wonder if Professor Reid's remark that STG "will likely be the most widely used" method of choosing λ might not be overly optimistic in these days of canned programs. Surely statisticians often use the transformation without any idea of using $\hat{\lambda}$ to point the way to a deeper consideration of what might be the correct scientific model, but for the other reasons offered by Professors Box and Cox—namely, to promote a linear model, with homoscedasticity and normality of the errors, and hence to justify the usual linear model inferences. The transformation is then just a mathematical extension of "let's take logs" or "let's take square roots," but now at least it is clear that a parameter in the original model has been estimated and should be accounted for.

There is another possibility worth considering. Suppose one uses the data to give the MLE of λ, but finds that, if anything, this value almost contradicts any preferred "scientific" value; this may happen because of a mixture, for example of effects, or other sources of confusion, and then the other criteria offered by Box and Cox (homoscedasticity and normality) would again be persuasive. In the case of our gasoline data, the only "sensible" value would appear to be $\lambda = 1$, and yet various factors (measurement error, autocorrelation, etc.) have combined to give $\hat{\lambda} = 1.47$. It then seems sensible to compare the results for at least these two values of λ. In Figure R-2, top panel, the model fits are all "reasonable" for the values of λ shown. In the figure are displayed the data together with fitted regressions transformed back to the original scale for the list $\Lambda = 0.0, 0.5, 1, 1.5, 2$ and 2.5. Of course, one could also argue that the plot should pass through the origin; in Figure R-2, bottom panel, the model fits are shown with this no-intercept restriction. (Recall that Professor Yang's analysis covers this possibility.) The fits again appear quite good, but we have the disconcerting result that now $\hat{\lambda} = 1.33$, with the 95% likelihood-ratio confidence interval from 1.08 to 1.79 (not including 1). More than ever this suggests that the whole model is not correct, because of the sources of error mentioned above. In the Box–Cox toxicity example, $\lambda = -1.0$ is chosen for interpretability, although the MLE gives $\lambda = -0.75$. If the ANOVA table using the MLE is constructed, the analysis is hardly changed.

In these more general circumstances, where there is no very strong reason after the event to choose a λ as though we had known it all along, we feel strongly that it should be admitted that the transformation is being used to make linear model theory more justified. Subsequent analysis should then take into account the fact that λ is part of the model and has been estimated; when the MLE is used, the analysis should be made using the terms which we provide to calculate standard errors. In addition, tests for normality of the errors should use the tests and tables given in our paper. Here, we agree with Professor Abadir's comments concerning the preference for confidence bands based on likelihood versus Wald, although, if the transformation is decided on scientific grounds, and the linear model is assured, the Wald interval after transformation would surely be the one to use. However, further study is needed on the implications of the STG method when the final λ, say $\tilde{\lambda}$, is not the MLE, but is chosen for a reason of convenience and yet has in some sense been estimated; how then should one correct for this procedure? We feel that our correction factors, now applied to the model using $\tilde{\lambda}$, should give approximate (possibly too small) standard errors for the parameter estimates, but this has not been investigated fully.

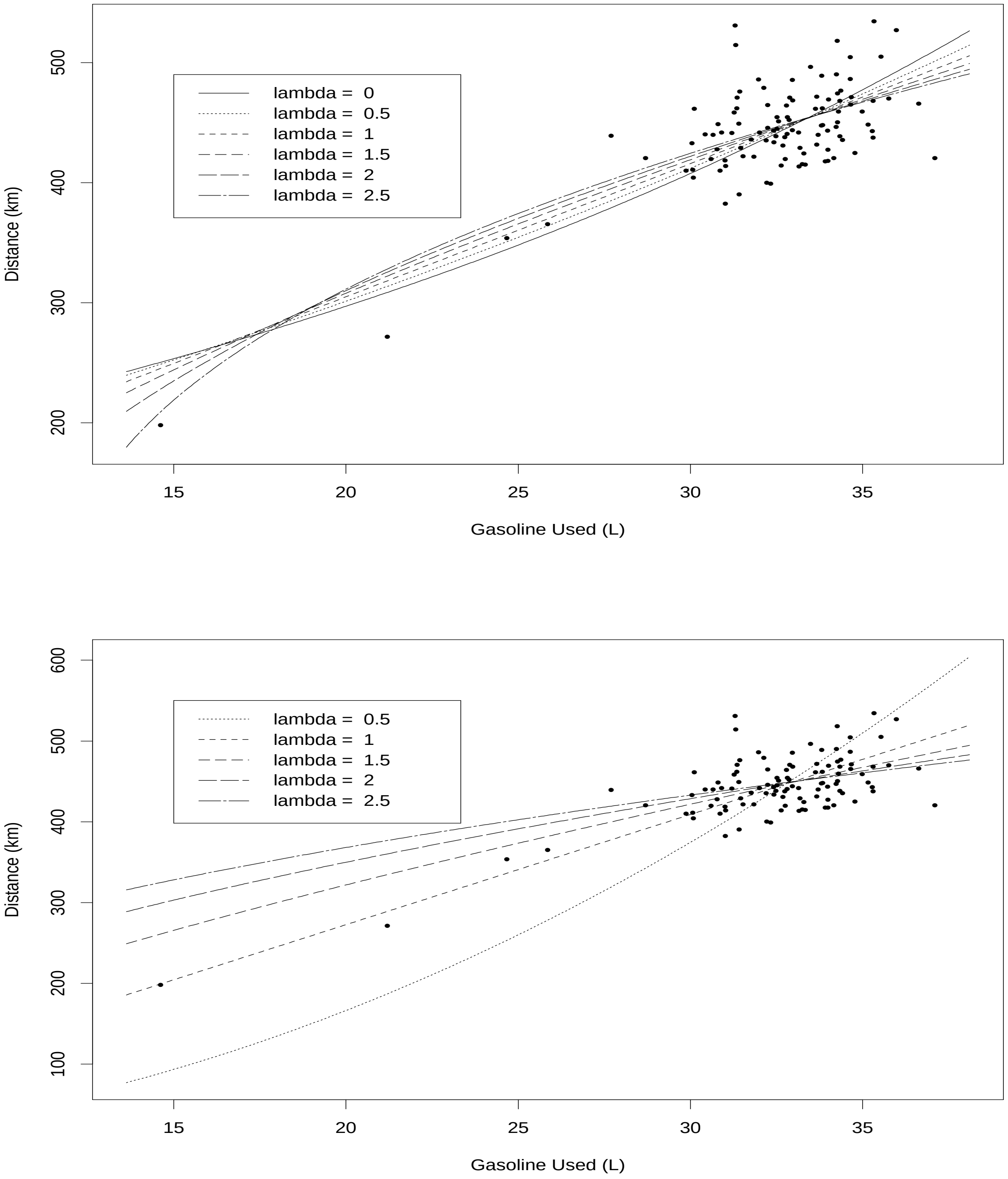

FIGURE R-2: Fitted curves on the original scale for the gasoline consumption example using $\lambda = 0.0$, 0.5, 1, 1.5, 2 and 2.5. Top panel: model as fitted in the paper. Bottom panel: model fitted without intercept term and without $\lambda = 0$.

3. *On the effect of non-normality*. In addition to the tests for normality, we have given correction factors when normality of errors cannot be assumed. For example, confidence bands for λ, given in Section 4.1, reduce to the classical form when errors are normal, but depend on g^* and g^{**} when they are not. As an illustration, suppose the errors have a Student t_p distribution, making them still symmetric, but longer-tailed than normal. The first six moments of the t_p distribution

are needed to calculate g^* and g^{**}, and then the revised confidence interval may be found from Z in Section 4.1, as suggested by Professor Abadir. These are shown in Figure 3 for several t_p error distributions. The intervals are longer than those based on the normal assumption, indicating that the true confidence level of the latter is smaller than nominal. Professors Reid and Bickel are justifiably perturbed that in practice one must *estimate* the six moments of the errors, and we agree that the attendant enormous variability is unfortunate. It would be interesting to make a Monte Carlo study to see how such estimation actually would affect, say, the confidence intervals for λ. For our data set, assuming a symmetric error distribution, and estimating μ_4 and $\mu + 6$ we find that the 95% interval using Z is (0.6,2.6) rather than the interval (0.7,2.4) given by assuming normality.

The above and similar reasons are why we wanted to give tests for the normal assumption. Our tests are omnibus tests; Professor Bickel's suggestion that one might tailor the tests to cover special alternatives might be covered by considering components of the given statistics (see Stephens 1986, Section 8.12); we look forward to the forthcoming paper by Professor Bickel and his co-authors giving tests with bootstrap critical points.

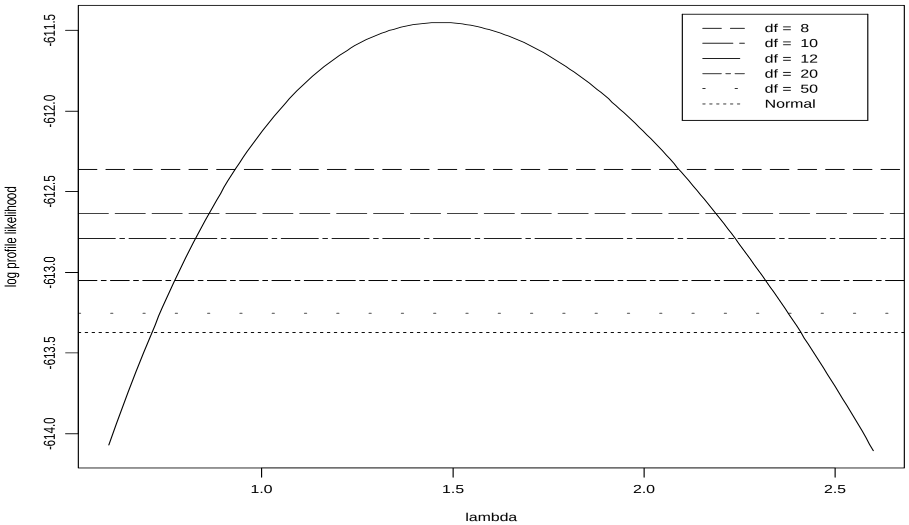

FIGURE R-3: Effect of heavy-tailed errors on likelihood ratio confidence intervals. The profile likelihood in Figure 1 of the paper is reproduced. The horizontal lines are the cut-offs using the adjustment based on statistic Z for t-distributed errors with the indicated degrees of freedom.

4. *On meanings and fortuitous properties*. Professor McCullagh provides convincing evidence that for response variables permitting aggregation (or "extensive variables" as they are called in Box & Cox 1964, bottom of p. 213) the form of the Box–Cox transformation is forced upon us (though applied only to the mean). The group-theoretic arguments apply exactly to the mean structure but only approximately to the error structure, and this error structure plays an important role in deriving the form of the estimators, and then again in the approximate distribution theory for these estimators. Once the approximate nature of the model has been acknowledged so plainly, it seems to us that an approximate analysis is called for. Algebraic arguments have an exact nature which diminishes their force considerably in approximate contexts.

Professor McCullagh challenges us to define "physically meaningful", and asks if we mean what he calls a "natural sub-parameter". "Physical" seems to us a fairly primitive concept and "meaningful" means "having meaning". As to "meaning" itself, we follow Good (1950, p. 1): "The meaning of 'meaning' will not be discussed". Over the years, statisticians have taken a number of useful English words and, by giving them precise mathematical meaning, abused their ordinary English meanings. Examples surely include "significant" and "bias". Perhaps "natural" is another example? The advantage of precise mathematical definitions is that we could now "prove" that θ is "natural". But we take note of Wittgenstein (1921): "A proposition of mathematics does not express a thought". Despite these caveats, we would regard "physically meaningful" as meaning what we understand Professor McCullagh to mean by "natural subparameter", at least in the contexts of our applications. Not only can the parameter, given in one set of units, be calculated in another, but in particular, the parameter, calculated on the transformed scale, can be compared with its value if it were known, or even estimated under the true λ scale—there is no danger of comparing apples with oranges, the kind of undesirable comparison rejected by Professors Box and Cox in their second paper in 1982.

We certainly agree that the pair (β, λ), and also (σ, λ) can be regarded as physically meaningful. It is possible to give a confidence band (using standard normal theory and assuming λ known) for these pairs, which might be useful in some circumstances. For our gasoline data set, these are shown in Figure R-4. Notice that, in principle, these confidence sets are infinite in extent in the λ direction. The sets emphasize that almost the entire uncertainty in, say, β is due to the uncertainty in λ.

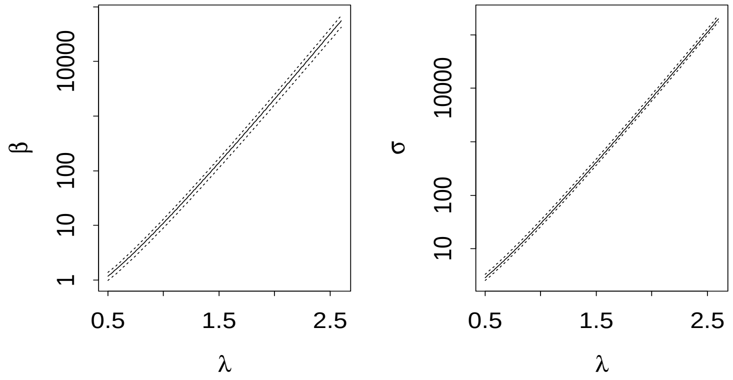

FIGURE R-4: Left: joint 95% confidence set for the pair (β, λ) derived from 95% confidence intervals for β using standard normal theory as if λ were known. The band is actually infinite in extent, running from $\lambda = -\infty$ to $\lambda = \infty$. Right: joint 95% confidence set for the pair (σ, λ) derived from 95% confidence intervals for σ using standard normal theory as if λ were known. The band is actually infinite in extent, running from $\lambda = -\infty$ to $\lambda = \infty$.

The work of Professor Johnson and his colleagues (Cho, Yeo, Johnson & Loh 2001a) provides a classical treatment of the asymptotics, including consistency, of the Box–Cox procedure making allowances for negative Y. Consistency is certainly an attractive property for many statisticians. In contrast, Professor McCullagh makes the observation that a property such as consistency or stability is largely fortuitous, and we agree—the consistency property, in general, might well be over-rated. As we emphasize in the paper, asymptotics are a guide to what will happen for real situations with a large sample, and, say, confidence bands for λ, as given in Section 4.1, are not less useful because an impossibly large sample would give a band which includes the "true" value (if such a value can be said to exist) but which does not necessarily become narrow.

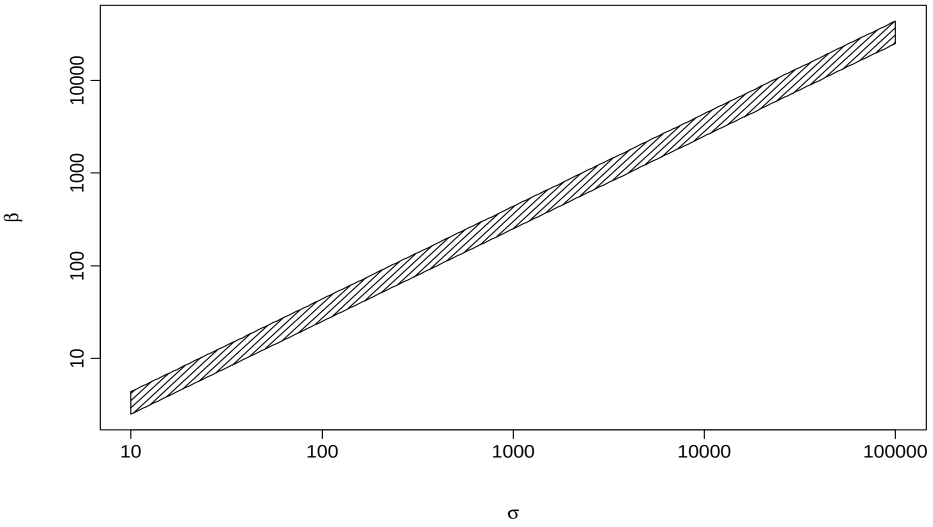

FIGURE R-5: Joint 95% confidence set for the pair (σ, β) derived from approximate 95% confidence interval for θ. The band is actually infinite in extent, running from $\sigma = 0$ to $\sigma = \infty$.

5. *The parameter θ.* Professor Bickel gives good reasons, other than stability, for considering θ as the relevant parameter, particularly if one considers a more general transformation than Box–Cox. However, Professor Hooper doubts the interpretability of θ in practice. It is true that θ is less informative if one regards β_i as giving the change in the (transformed) response caused by a unit change in $x_{\cdot i}$, as of course is reasonable when one is sure that the $x_{\cdot i}$'s in multiple regression are not highly correlated. However, consider the case of random covariates, i.e., where $x_1, \ldots, x_n$ are a sample from some population (univariate for simplicity of discussion). If the relation between the y_i and x_i is a simple linear regression, then a common summary statistic is the sample Pearson correlation coefficient r. In the Box–Cox context, our model (2) would specify, approximately, the conditional distribution of the responses y_i given the covariates. For a given λ, we could then consider the behaviour of the fitted correlation r between the values of y_i transformed using λ and the x_i. It is elementary that

$$r = \hat{\theta} s_x / \sqrt{1 + (\hat{\theta} s_x)^2}\,,$$

where now s_x is the sample standard deviation of the x_i values, but where both $\hat{\sigma}$ (used in $\hat{\theta}$) and s_x have been calculated by dividing by n (and not $n-p$ and $n-1$, respectively). The calculations in our paper then give, by Taylor expansion, standard errors for the true correlation conditional on the x_i values. Unconditional standard errors may also be derived by further expansion but will depend, in general, on the second and fourth moments of the x population. In Figure R-6, we show that r is reasonably stable.

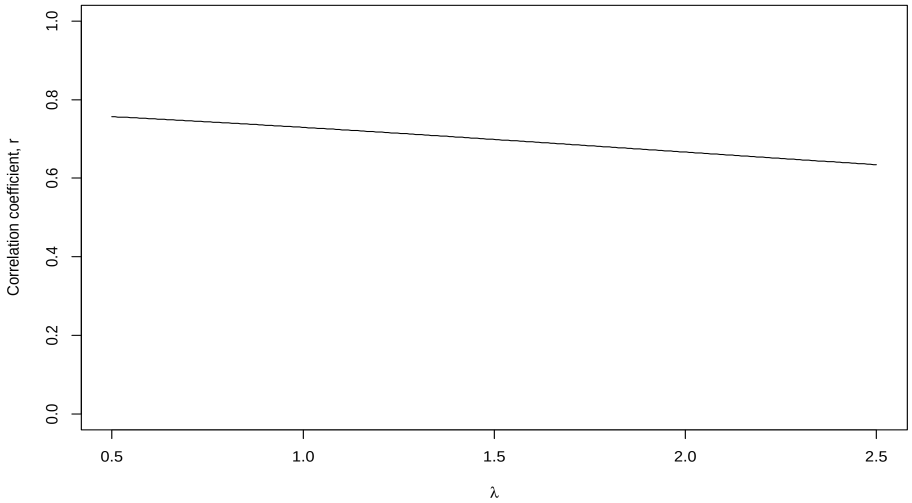

FIGURE R-6: Plot of the Pearson correlation r against λ for the gasoline data in the paper.

When the covariate values are determined by the experimenter (as is actually the case for our gasoline data), the experimenter can control r by controlling s_x. The standardized slope, however, cannot be so controlled. In view of the relation

$$\hat{\theta} = \frac{r}{s_x\sqrt{1-r^2}} \; ,$$

statistic $\hat{\theta}$ may have potential as a replacement for r in designed experiments.

6. *The parameter β*. However, if it is β that the applied worker wants inference about, as Professor Hooper challenges, then we have to say that it is fortuitously a pity that the estimate of this parameter is unstable—but there it is. We remind ourselves that our first goal was to give a large-sample model which would describe the plots in Figure 1 of the paper, plots which arose from real data, whether or not we like them!

Professor Doksum draws our attention to the stability of $\hat{\beta}/|\hat{\beta}|$, and we are grateful to note the connections he makes to other work. When $p = 1$ (as for the gasoline data), this quantity is *extremely* stable; its value is 1 for all λ. Of course it is intended for use only for $p \geq 2$ but there, as Professor Doksum notes, it is just one summary of $\hat{\theta}$ since $\hat{\theta}/|\hat{\theta}| = \hat{\beta}/|\hat{\beta}|$. Again, large-sample theory is available for this quantity away from $\theta = 0$ in our framework. Professor Doksum's tables and graphs show that this normalized θ is more stable than θ itself, but at the price of losing information about each component. This would seem to be a serious drawback.

7. *Other model formulations*. We now make some brief remarks on other possible models.

Professor Reid asks if use of the adjusted likelihood $\ell_{CR}(\lambda)$ might make the estimate of λ "more nearly consistent". As she hints, to the order of approximation studied here, and under our conditions, maximization of ℓ_{CR} leads to the same normal approximations for the estimator. It does seem useful to us to ask what higher order expansions, retaining some terms in δ for instance—as does Professor Yang—might tell us. We remark that our calculations permit p to grow somewhat with n and that our earlier draft of this paper showed that our results remain valid when $|\theta|$ grows with n.

Professor Hooper suggests consideration of the parameter function

$$\beta(\lambda) = \mathrm{E}_{\lambda_0}\{\hat{\beta}(\lambda)\} \equiv \psi(\lambda, \lambda_0, \mu_0, \beta_0, \sigma_0, X),$$

where we use the function ψ simply to emphasize the dependence, under the model, on both the true values of the parameters λ, β, and σ (denoted by the subscript 0) and on an assumed value of λ, as well as on the current design points X. He would treat $\beta(\hat{\lambda})$ as a parameter and get confidence intervals for $\psi(\hat{\lambda}, \lambda_0, \mu_0, \beta_0, \sigma_0, X)$ from the distribution of $\hat{\beta} - \beta(\hat{\lambda})$, where $\hat{\beta} = \hat{\beta}(\hat{\lambda})$ is the MLE of β as used in our paper. Professor Hooper is in good company in considering a data-dependent parameter; the Box–Cox device of dividing by the geometric mean of the y values effectively makes $\beta(\hat{\lambda})$ data-dependent. We have some concern about the interpretation of this "parameter", but note that our asymptotic framework can provide some insight into the behaviour of the procedure. Put $\phi = \delta(\lambda - \lambda_0)/\lambda_0$, corresponding to ϕ in our paper, and set $\phi/\delta = w$. Then

$$\hat{\beta}(\phi) - \beta(\phi) = \hat{\beta}(\phi) - \mathrm{E}_{\lambda_0}\{\hat{\beta}(\phi)\} = (1 + \lambda_0\mu_0)^w \left[\tilde{\beta}(\phi) - \mathrm{E}_{\lambda_0}\{\tilde{\beta}(\phi)\}\right].$$

The expansions in our Theorem 2 give approximately (passing expectations through expansions)

$$\mathrm{E}_{\lambda_0}\{\tilde{\beta}(\phi)\} = \beta + \sigma(X_1'X_1)^{-1}X_1'\eta^{\#2}\phi = \beta + \sigma v\phi/2$$

in our notation. Professor Hooper is thus estimating roughly

$$(1 + \lambda_0\mu_0)^w(\beta_0 + \sigma_0 v\phi/2).$$

We find that

$$\hat{\beta} - \beta(\hat{\lambda}) = (1 + \lambda_0\mu_0)^w\sigma_0(X_1'X_1)^{-1}X_1'\varepsilon$$

approximately. If $\hat{w} = \hat{\phi}/\delta = (\hat{\lambda} - \lambda_0)/\lambda_0 = o_P(1)$, then, as Professor Hooper says, the asymptotic variance of $\hat{\beta} - \beta(\hat{\lambda})$ is $\sigma_0^2(X_1'X_1)^{-1}$, i.e., the usual variance of ordinary least squares; there is no variance inflation. If, however, $\delta\sqrt{g}/\lambda_0$ does not tend to infinity (and it need not under our conditions), then the factor $(1 + \lambda_0\mu_0)^w$ inflates the variability of $\hat{\beta} - \beta(\hat{\lambda})$ and also makes the distribution of this difference *not* normal.

Fortunately, however, confidence intervals for $a'\beta(\hat{\lambda})$ are based on the distribution of

$$\frac{a'\{\hat{\beta} - \beta(\hat{\lambda})\}}{\hat{\sigma}\sqrt{a'(X_1'X_1)^{-1}a}}$$

which, under our set-up converges in law to $\mathsf{N}(0, 1)$. Thus, even in our framework, and in spite of the variation inflation in $\hat{\beta} - \beta(\hat{\lambda})$, standard regression intervals are valid (even for β near 0).

If $\beta(\hat{\lambda})$ is a valid estimand, we would expect that tests of $\mathcal{H}_0 : a'\beta(\hat{\lambda}) = 0$ would make sense. But this would appear to test, approximately, the hypothesis that

$$a'\beta_0 + \delta\sigma_0 a'v(\hat{\lambda} - \lambda)/(2\lambda_0) = 0$$

or equivalently

$$a'\theta_0 + \frac{\delta\sigma_0 a'v(\hat{\lambda} - \lambda)}{2(1 + \lambda_0\mu_0)} = 0.$$

We find this hypothesis difficult to interpret since $a'\theta_0 = 0$ does not guarantee $a'v = 0$.

Professor Yang's model, as he says, is closely related to ours; one difference is that he puts the intercept μ in with the usual β's, so that his formulas are slightly different from ours. However, the important quantity g, which first appeared in the doctoral dissertation of the first author (Chen 1991) is parallelled in the denominator of his expansions for $(\lambda - \lambda_0)/\lambda_0$. As our paper came to fruition we were not aware of Yang (1999), where the model formulation is similar to his discussion. Another related paper, Yang (2002), appears also in this issue of *The Canadian Journal of Statistics*.

We have appreciated all the suggestions made on alternative ways to examine the Box–Cox transformation which have made the discussion so interesting. It would be interesting to investigate and compare these alternatives, particularly as applied to specific data sets. However, this would take weeks if not months, so for now we merely point out that the fundamental paper of Professors Box and Cox, and all the possibilities stemming from it, will surely provide grist to the Ph. D. mill for decades to come.

ADDITIONAL REFERENCES USED IN THE DISCUSSION

K. M. Abadir (1999). An introduction to hypergeometric functions for economists. *Econometric Reviews*, 18, 287–330.

A. C. Atkinson (1985). *Plots, Transformations, and Regression: an Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press, New York.

P. J. Bickel & Y. Ritov (1997). LAN of ranks and covariates in the transformation models. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen & G. Yang, eds). Springer-Verlag, New York, pp. ???–???.

P. J. Bickel, Y. Ritov & ?. Ryden (200?). Submitted for publication in *The Annals of Statistics*.

P. J. Bickel, Y. Ritov & T. Stoker (2001). Tailor made tests of goodness-of-fit for semiparametric hypotheses. Mimeo.

D. R. Brillinger (1983). A generalized linear model with "Gaussian" regressor variables. *Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum & J. L. Hodges, Jr., eds), Wadsworth, Belmont, CA, pp. 97–114.

R. J. Carroll & D. Ruppert (1981). Prediction and the power transformation family. *Biometrika*, 68, 609–616.

E. A. Catchpole & B. J. T. Morgan (1997). Detecting parameter redundancy. *Biometrika*, 84, 187–196.

P. Chaudhuri, K. A. Doksum & A. Samarov (1997). On average derivative quantile regression. *The Annals of Statistics*, 25, 715–744.

G. Chen (1991). *Empirical Process Based on Regression Residuals: Theory and Applications*. Unpublished doctoral dissertation, Department of Mathematics and Statistics, Simon Fraser University, Burnaby, British Columbia, Canada.

K. Cho, I. Yeo, R. A. Johnson & W.-Y. Loh (2001a). Asymptotic theory for Box–Cox transformations in linear models. *Statistics & Probability Letters*, 51, 337–343.

K. Cho, I. Yeo, R. A. Johnson & W.-Y. Loh (2001b). Prediction interval estimation in transformed linear models. *Statistics & Probability Letters*, 51, 345–350.

A. Cohen & H. B. Sackrowitz (1987). An approach to inference following model selection with applications to transformation-based and adaptive inference. *Journal of the American Statistical Association*, 82, 1123–1130.

F. Critchley, P. Marriott & M. Salmon (1996). On the differential geometry of the Wald test with nonlinear restrictions. *Econometrica*, 64, 1213–1222.

M. Dagenais & J.-M. Dufour (1991). Invariance, nonlinear models, and asymptotic tests. *Econometrica*, 59, 1601–1615.

R. B. Davies (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64, 247–254.

K. Doksum (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *The Annals of Statistics*, 15, 325–345.

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

I. J. Good (1950). *Probability and the Weighing of Evidence*. Griffin, London.

W. Härdle & T. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal American Statistical Association*, 84, 986–995.

?. Kadane, P. J. Bickel & K. Doksum (1989).

S. D. C. Lawford (2001). *Improved Modelling in Finite-Sample and Nonlinear Frameworks*. Unpublished D. Phil. thesis, University of York, United Kingdom.

K.-C. Li (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the Amer-*

ican Statistical Association, 86, 316–342.

P. McCullagh (2002). What is a statistical model? (with discussion). *The Annals of Statistics*, 30, in press.

T. J. Rothenberg (1971). Identification in parametric models. *Econometrica*, 39, 577–591.

M. A. Stephens (1975). Asymptotic properties of covariance matrices of order statistics. *Biometrika*, 62, 23–28.

M. A. Stephens (1986). Tests for the exponential distribution. In *Goodness-of-Fit Techniques* (R. B. D'Agostino & M. A. Stephens, eds). Marcel Dekker, New York, pp. 421–460.

T. Stoker (1986). Consistent estimation of scaled coefficients. *Econometrica*, 54, 1461–1481.

J. M. G. Taylor (19899).

L. Wittgenstein (1921). *Tractatus Logico-Philosophicus.* Translated, 1961, by D. F. Pears & B. F. McGuinness. Routledge, London.

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

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

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