

The Box and Cox Problem: Asymptotic Significance Levels

D.A.S. Fraser

Department of Statistics
University of Toronto
Toronto, Canada, M5S 3G3

X. Yuan

Department of Mathematics and Statistics
Okanagan University College
Kelowna, B.C., Canada, V1V 1V7

SUMMARY

The common linear model describes responses y_1, \dots, y_n that are independent and normally distributed with constant variance and means specified by a linear function of regression parameters β . Box & Cox (1964) considered the more general situation where the initial response y is transformed to $y^{(\psi)} = h(\psi; y)$ for some ψ and produces a modified response $y^{(\psi)}$ with the preceding linear model. They used likelihood and Bayesian methods for inference concerning the transformation parameter ψ . We apply recent asymptotic theory to obtain higher order significance levels for testing the scalar parameter ψ ; simulations provide an assessment of the accuracy of the approximate significance level. Generalizations to transformations of the input variables are discussed.

Keywords: likelihood asymptotics; power transformation; transformed data

1. INTRODUCTION

A familiar problem in science is that of finding an appropriate mode of expression for a response variable so that its behaviour relative to input variables is simple and easily understood. Box & Cox (1964) examined transformations of some initial response to obtain a transformed variable satisfying a linear model with say additive effects for independent factors, constancy of error variance, and perhaps normality. They gave a detailed discussion of these issues from a statistical and scientific viewpoint.

Consider some initial response variable y and a transformed response $y^{(\psi)} = h(\psi; y)$ that is a monotone increasing function of y . A familiar transformation in applications is the power transformation y^ψ on the positive axis. For negative ψ this becomes monotone decreasing; accordingly the modification

$$\begin{aligned} h(\psi; y) &= \frac{y^\psi - 1}{\psi} & \psi \neq 0 \\ &= \log y & \psi = 0 \end{aligned} \tag{1.1}$$

is frequently used. This modification is monotone and converges to $\log y$ as $\psi \rightarrow 0$. Box & Cox (1964) also considered a change of origin for the initial y . In this paper we restrict attention to monotone increasing transformations from y to $y^{(\psi)}$ with a scalar ψ . The methods can be extended to vector ψ by examining one ψ coordinate at a time, and to nonmonotone transformations.

Consider a vector of observations $y = (y_1, \dots, y_n)'$ such that the corresponding transformed vector $y^{(\psi)} = (y_1^{(\psi)}, \dots, y_n^{(\psi)})'$ for some particular value of ψ satisfies a linear model $X\beta + \sigma e$ where e is a vector of independent and identically distributed errors and X is an $n \times r$ full column-rank design matrix; thus the full parameter is $\theta = (\beta', \sigma, \psi)'$. We focus on the case with standard normal errors e but the case with other error form, say Student (6), is also discussed but can be somewhat more difficult computationally.

Box & Cox (1964) develop a profile likelihood analysis and a Bayesian analysis based on a data dependent prior; the Bayesian analysis corrects what seems to be an anomalous

degrees-of-freedom that arises in the likelihood analysis. The full log likelihood function is

$$\ell(\theta) = \ell(\beta, \sigma, \psi) = a - \frac{1}{2\sigma^2} \left| y^{(\psi)} - X\beta \right|^2 - \frac{n}{2} \log \sigma^2 + \log J(\psi; y) \quad (1.2)$$

where $J(\psi; y) = \prod dy_i^{(\psi)} / dy_i$ is the Jacobian of the transformation to the new response.

The profile likelihood $\ell_p(\psi)$ for ψ is obtained by maximizing over β, σ ,

$$\ell_p(\psi) = \ell(\hat{\theta}_\psi; y) = \ell(\hat{\beta}_\psi, \hat{\sigma}_\psi, \psi) = a - \frac{n}{2} \log \hat{\sigma}^2(\psi) + \log J(\psi, y) = a - \frac{n}{2} \log \{\hat{\sigma}^2(\psi; z)\} , \quad (1.3)$$

where $\hat{\sigma}^2(\psi; y) = S^2(\psi; y)/n$ with $S^2(\psi; y)$ equal to the sum of squares of residuals for $y^{(\psi)}$ and $\hat{\sigma}^2(\psi; z) = S^2(\psi; z)/n$ with $S^2(\psi; z)$ equal to the sum of squares of residuals for the rescaled response vector

$$z^{(\psi)} = y^{(\psi)} / J^{1/n}(\psi; y) . \quad (1.4)$$

The maximum likelihood value for ψ can be obtained by minimizing $S^2(\psi; z)$. The Bayesian analysis leads to an expression similar to (1.3) but with the usual degrees of freedom $n - r$ replacing n in the definitions of the variance estimates; the modified expression is called the Bayesian adjusted profile likelihood.

First order asymptotic theory leads to a chi square (1) distribution for

$$r_\psi^2 = 2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\} \quad (1.5)$$

or a standard normal distribution for the signed likelihood ratio

$$r_\psi = \text{sgn}(\hat{\psi} - \psi) \cdot [2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}]^{1/2} . \quad (1.6)$$

The corresponding first order significance function for assessing ψ is then $p_1(\psi) = \Phi(r_\psi)$ where Φ is the standard normal distribution function and observed values have been substituted; this is a p -value recording probability nominally left of the observed $\hat{\psi}^0$ when the parameter has the value ψ .

In this paper we use recent asymptotic theory to obtain a third order significance function $p(\psi)$ for assessing ψ . Section 2 gives a brief outline of the needed third order

asymptotic theory. Section 3 discusses this for the present formulation of the Box & Cox problem, and Section 4 discusses two examples, one with real data and one simulated. Section 5 records a Monte Carlo study to assess the accuracy of the third order significance levels. Section 6 gives a brief overview and literature review.

2. THIRD ORDER ASYMPTOTIC METHODS

Third order inference for a scalar parameter component $\psi = \psi(\theta)$ in a continuous statistical model $f(y; \theta)$ was developed by Barndorff-Nielsen (1986) for the case where the full variable and parameter were of the same dimension say p . The departure measure was a modified likelihood ratio that was computationally difficult but a more accessible variant was obtained as a mean and scale corrected likelihood ratio. A modified general version was obtained in Barndorff-Nielsen (1991).

An alternative third order approach using tangent exponential models (Fraser, 1990) was developed (Fraser & Reid, 1990, 1993, 1995) for the same dimension case. This used the ordinary likelihood ratio r and a standardized maximum likelihood departure derived from the tangent model; these were then combined to give the significance function $p(\psi)$ by one or other of the Lugannani & Rice (1980) formula

$$p(\psi) = \Phi(r) + \varphi(r)(r^{-1} - q^{-1}) \quad (2.1)$$

or the Barndorff-Nielsen (1991) formula

$$p(\psi) = \Phi\{r - r^{-1} \log(r/q)\} ; \quad (2.2)$$

for this $\varphi(z)$, $\Phi(z)$ are the standard normal density and distribution functions and $p(\psi)$ records probability left of the data in the nominal sense $\hat{\psi} \leq \hat{\psi}^0$ when the parameter has value ψ . Most third order significance formulas can be presented in terms of these combining formulas with r as the signed likelihood root

$$r = \text{sgn}(\hat{\psi} - \psi)[2\{\ell(\hat{\theta}; y) - \ell(\hat{\theta}_\psi; y)\}]^{1/2} \quad (2.3)$$

and q given by some expression appropriate to the particular problem; in (2.3), $\hat{\theta}$ designates the overall maximum likelihood value and $\hat{\theta}_\psi$ the constrained maximum likelihood value given $\psi(\theta) = \psi$.

For the tangent exponential model approach the q is taken to be a standardized maximum likelihood departure

$$q = (\hat{\chi} - \hat{\chi}_\psi) \frac{|\hat{j}_{(\theta\theta)}|^{1/2}}{|j_{(\lambda\lambda)}(\hat{\theta}_\psi)|^{1/2}}. \quad (2.4)$$

This uses a nominal reparameterization

$$\varphi(\theta) = \frac{\partial}{\partial \mathbf{y}} \ell(\theta; \mathbf{y}) \Big|_{\mathbf{y}^0} = \ell_{, \mathbf{y}}(\theta; \mathbf{y}^0) \quad (2.5)$$

based on the tangent exponential model which is then used to construct a replacement

$$\chi(\theta) = \frac{\psi_{\varphi'}(\hat{\theta}_\psi^0)}{|\psi_{\varphi'}(\hat{\theta}_\psi^0)|} \varphi(\theta) \quad (2.6)$$

for $\psi(\theta)$, where the gradient of $\psi(\theta)$

$$\psi_{\varphi'}(\hat{\theta}_\psi^0) = \frac{\partial}{\partial \varphi'} \psi(\theta) \Big|_{\hat{\theta}_\psi^0} \quad (2.7)$$

makes $\chi(\theta)$ mimic key properties of $\psi(\theta)$ near $\hat{\theta}_\psi$. Also the information $\hat{j}_{(\theta\theta)}$ is the full parameter information recalibrated in terms of φ and $j_{(\lambda\lambda)}(\hat{\theta}_\psi)$ is the nuisance parameter information recalibrated also in terms of φ where the parameter is separated as $\theta = (\lambda', \psi)'$.

A sufficiency reduction as with an exponential model or a conditioning reduction as with a transformation model may lead to the direct application of the preceding methods. More generally an approximate conditioning reduction may lead effectively to the same dimension case. A method for obtaining such a reduction is developed in Fraser & Reid (1995). For the use of (2.1) or (2.2) with (2.3) and (2.4) only the tangent directions $V = (v_1, \dots, v_p)$ to the conditioning surface at the observed data point \mathbf{y}^0 are needed. Tangent location model theory (Fraser, 1964) leads easily (Fraser & Reid, 1995) to such

tangent directions. For the case of independent coordinates with distribution function $F_i(y_i; \theta)$, we obtain the $n \times p$ array

$$V = \left(-\frac{\partial F_1(y_1; \theta)/\partial \theta}{\partial F_1(y_1; \theta)/\partial y_1}, \dots, -\frac{\partial F_n(y_n; \theta)/\partial \theta}{\partial F_n(y_n; \theta)/\partial y_n} \right)' \Big|_{(y^0, \hat{\theta}^0)}. \quad (2.8)$$

It then suffices to take the nominal reparameterization $\varphi(\theta)$ to be the gradient

$$\varphi'(\theta) = \frac{d}{dV} \ell(\theta; y) \Big|_{y^0} = \ell_{;V}(\theta; y^0) \quad (2.9)$$

of the log likelihood in the conditioning directions.

3. SIGNIFICANCE FOR THE BOX AND COX PROBLEM

Consider some initial response y and a transformation $y^{(\psi)} = h(\psi; y)$ that is monotone increasing in y for each value of a scalar parameter ψ . We consider a vector $y = (y_1, \dots, y_n)$ of observations and assume that for some ψ the transformed response vector $y^{(\psi)} = (y_1^{(\psi)}, \dots, y_n^{(\psi)})$ has a linear model

$$y^{(\psi)} = X\beta + \sigma e \quad (3.1)$$

where X is of full column rank r and e is a sample from an error distribution $f(e)$ which could be say standard normal or rescaled Student (6) or other. We are concerned with inference for the parameter ψ .

The full log likelihood $\ell(\theta; y) = \sum_1^n \ell_i(\theta; y_i)$ is a sum of contributions,

$$\ell_i(\theta; y_i) = -\frac{1}{2} \log \sigma^2 + g\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\} + \log h_{;y}(\psi; y_i), \quad (3.2)$$

from the component observations where X_i is the i th row of X , $g(e) = \log f(e)$, and $h_{;y}(\psi; y) = \partial h(\psi; y)/\partial y$. The score quantities are

$$\begin{aligned} \ell_\beta(\theta; y) &= -\frac{1}{\sigma} \sum_1^n g'\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\} X_i' \\ \ell_\sigma(\theta; y) &= -\frac{n}{\sigma} - \frac{1}{\sigma^2} \sum_1^n g'\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\} (y_i^{(\psi)} - X_i\beta) \\ \ell_\psi(\theta; y) &= \frac{1}{\sigma} \sum_1^n g'\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\} h_\psi(\psi; y_i) + \sum_1^n h_{\psi;y}(\psi; y_i) h_{;y}^{-1}(\psi; y_i) \end{aligned} \quad (3.3)$$

where $h_\psi(\psi, y) = \partial h(\psi; y)/\partial \psi$ and $h_{\psi; y}(\psi; y) = \partial^2 h(\psi; y)/\partial \psi \partial y$. For the case of normal error $g'(e) = -e$, and for the power transformation case (1.1) $h_{; y}(\psi; y) = y^{\psi-1}$, $h_\psi(\psi; y) = \psi^{-1}(y^\psi \log y - h(\psi; y))$, and $h_{\psi; y}(\psi; y) = \log y \cdot y^{\psi-1}$.

The overall maximum likelihood value $\hat{\theta} = (\hat{\beta}', \hat{\sigma}, \hat{\psi})' = \hat{\theta}(y)$ can be obtained by iteratively solving $\ell_\theta = 0$. The constrained maximum likelihood value $\hat{\theta}_\psi = (\hat{\beta}'_\psi, \hat{\sigma}_\psi, \psi)' = \hat{\theta}_\psi(y)$ can be obtained by iteratively solving $\ell_\beta = 0$, $\ell_\sigma = 0$ with ψ fixed. In the normal case

$$\begin{aligned}\hat{\beta}_\psi &= (X'X)^{-1}X'y^{(\psi)} \\ \hat{\sigma}_\psi^2 &= \{y'^{(\psi)}y^{(\psi)} - y'^{(\psi)}X(X'X)^{-1}X'y^{(\psi)}\}/n\end{aligned}\tag{3.4}$$

and the profile likelihood for ψ is then given by (1.3).

The information components can be calculated from the second derivatives of likelihood:

$$\begin{aligned}\ell_{\beta\beta'}(\theta; y) &= \frac{1}{\sigma^2} \sum_1^n g''\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\}X'_iX_i \\ \ell_{\beta\sigma}(\theta; y) &= \frac{1}{\sigma^2} \sum_1^n g'\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\}X'_i + \frac{1}{\sigma^3} \sum_1^n g''\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\}X'_i\{y_i^{(\psi)} - X_i\beta\} \\ \ell_{\sigma\sigma}(\theta; y) &= \frac{n}{\sigma^2} + \frac{2}{\sigma^3} \sum_1^n g'\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\}\{y_i^{(\psi)} - X_i\beta\} \\ &\quad + \frac{1}{\sigma^4} \sum_1^n g''\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\}\{y_i^{(\psi)} - X_i\beta\}^2 \\ \ell_{\beta\psi}(\theta; y) &= -\frac{1}{\sigma^2} \sum_1^n g''\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\}X'_ih_\psi(\psi; y_i) \\ \ell_{\sigma\psi}(\theta; y) &= -\frac{1}{\sigma^3} \sum_1^n g''\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\}\{y_i^{(\psi)} - X_i\beta\}h_\psi(\psi; y_i) \\ &\quad - \frac{1}{\sigma^2} \sum_1^n g'\{\sigma^{-1}\{y_i^{(\psi)} - X_i\beta\}\}h_\psi(\psi; y_i) \\ \ell_{\psi\psi}(\theta; y) &= \frac{1}{\sigma^2} \sum_1^n g''\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\}h_\psi^2(\psi; y_i) + \frac{1}{\sigma} \sum_1^n g'\{\sigma^{-1}(y_i^{(\psi)} - X_i\beta)\}h_{\psi\psi}(\psi; y_i) \\ &\quad + \sum_1^n h_{\psi\psi; y}(\psi; y_i)h_{; y}^{-1}(\psi; y_i) - \sum_1^n h_{\psi; y}^2(\psi; y_i)h_{; y}^{-2}(\psi; y_i)\end{aligned}\tag{3.5}$$

where the subscripts to $h(\psi; y)$ denote differentiation with respect to ψ and to y . The overall information matrix is

$$\hat{j}_{\theta\theta} = -\ell_{\theta\theta'}(\hat{\theta}; y) \quad (3.6)$$

and the constrained information matrix is

$$j_{\lambda\lambda}(\hat{\theta}_\psi) = -\ell_{\lambda\lambda'}(\hat{\theta}_\psi; y) \quad (3.7)$$

where $\lambda = (\beta', \sigma)'$.

For the normal case the full information is

$$\hat{j}_{\theta\theta} = \begin{bmatrix} \hat{\sigma}^{-2} X'X & 0 & -\hat{\sigma}^{-2} \sum_1^n X_i' h_\psi(\hat{\psi}; y_i) \\ 0 & 2n\hat{\sigma}^{-2} & -2\hat{\sigma}^{-3} \sum_1^n \{y_i^{(\hat{\psi})} - X_i\hat{\beta}\} h_\psi(\hat{\psi}; y_i) \\ - & - & \hat{j}_{\psi\psi} \end{bmatrix}, \quad (3.8)$$

where

$$\hat{j}_{\psi\psi} = \hat{\sigma}^{-2} \sum_1^n [h_\psi^2(\hat{\psi}; y_i) + \{y_i^{(\hat{\psi})} - X_i\hat{\beta}\} h_{\psi\psi}(\hat{\psi}; y_i)],$$

and the nuisance parameter information is

$$j_{\lambda\lambda}(\hat{\theta}_\psi) = \begin{bmatrix} \hat{\sigma}_\psi^{-2} X'X & 0 \\ 0 & 2n\hat{\sigma}_\psi^{-2} \end{bmatrix}. \quad (3.9)$$

Consider now the nominal reparameterization $\varphi(\theta)$. For this the gradient $m_i(\theta; y_i) = d\ell_i(\theta; y_i)/dy_i$ of the i th component of likelihood is

$$m_i(\theta; y_i) = \sigma^{-1} g'[\sigma^{-1} \{y_i^{(\psi)} - X_i\beta\}] h_{;y}(\psi; y_i) + h_{;yy}(\psi; y_i) h_{;y}^{-1}(\psi; y_i), \quad (3.10)$$

which in the standard normal case with the power transformation $y^{(\psi)} = (y^\psi - 1)/\psi$ becomes

$$m_i(\theta; y_i) = -\sigma^{-2} \{y_i^{(\psi)} - X_i\beta\} y_i^{\psi-1} + y_i^{-1}(\psi - 1). \quad (3.11)$$

The tangent directions can be obtained from the coordinate distribution functions or more easily from the coordinate pivotal $\{y_i^{(\psi)} - X_i\beta\}/\sigma$. Setting the total differential of the coordinate pivotal equal to zero,

$$\sigma^{-1} h_{;y}(\psi; y_i) dy_i - \sigma^{-1} X_i d\beta - \sigma^{-2} \{y_i^{(\psi)} - X_i\beta\} d\sigma + \sigma^{-1} h_{\psi}(\psi; y_i) d\psi = 0,$$

gives

$$dy_i = h_{,y}^{-1}(\psi; y_i) X_i d\beta + \sigma^{-1} h_{,y}^{-1}(\psi; y_i) \{y_i^{(\psi)} - X_i \beta\} d\sigma - h_{,y}^{-1}(\psi; y_i) h_{\psi}(\psi; y_i) d\psi \quad (3.12)$$

from which we obtain the reparameterization $\varphi(\theta)$ with coordinates

$$\begin{aligned} \varphi_1 &= \sum_1^n m_i(\theta; y_i) h_{,y}^{-1}(\hat{\psi}; y_i) X_i' \\ \varphi_2 &= \sum_1^n m_i(\theta; y_i) h_{,y}^{-1}(\hat{\psi}; y_i) \{y_i^{(\hat{\psi})} - X_i \hat{\beta}\} \hat{\sigma}^{-1} \\ \varphi_3 &= - \sum_1^n m_i(\theta; y_i) h_{,y}^{-1}(\hat{\psi}; y_i) h_{\psi}(\hat{\psi}; y_i) . \end{aligned} \quad (3.13)$$

For the case of standard normal error and the power transformation $y(\psi; y) = (y^{\psi} - 1)/\psi$, the reparameterization simplifies to

$$\begin{aligned} \varphi_1 &= \sum_1^n m_i(\theta; y_i) y_i^{1-\hat{\psi}} X_i' \\ \varphi_2 &= \sum_1^n m_i(\theta; y_i) y_i^{1-\hat{\psi}} \{y_i^{(\hat{\psi})} - X_i \hat{\beta}\} \hat{\sigma}^{-1} \\ \varphi_3 &= \sum_1^n m_i(\theta; y_i) y_i^{1-\hat{\psi}} \hat{\psi}^{-1} \{y_i^{(\hat{\psi})} - y_i^{\hat{\psi}} \log y_i\} . \end{aligned} \quad (3.14)$$

For the Jacobian of the transformation from θ to φ we first calculate

$$M_i(\theta; y_i) = \left\{ \frac{\partial m_i(\theta; y_i)}{\partial \beta'}, \frac{\partial m_i(\theta; y_i)}{\partial \sigma}, \frac{\partial m_i(\theta; y_i)}{\partial \psi} \right\} , \quad (3.15)$$

where

$$\begin{aligned} \frac{\partial m_i(\theta; y_i)}{\partial \beta'} &= -\sigma^{-2} g''[\sigma^{-1}\{y_i^{(\psi)} - X_i \beta\}] X_i h_{,y}(\psi; y_i) \\ \frac{\partial m_i(\theta; y_i)}{\partial \sigma} &= -\sigma^{-2} g'[\sigma^{-1}\{y_i^{(\psi)} - X_i \beta\}] h_{,y}(\psi; y_i) \\ &\quad - \sigma^{-3} g''[\sigma^{-1}\{y_i^{(\psi)} - X_i \beta\}] h_{,y}(\psi; y_i) \{y_i^{(\psi)} - X_i \beta\} \\ \frac{\partial m_i(\theta; y_i)}{\partial \psi} &= \sigma^{-1} g'[\sigma^{-1}\{y_i^{(\psi)} - X_i \beta\}] h_{\psi; y}(\psi; y_i) \\ &\quad + \sigma^{-2} g''[\sigma^{-1}\{y_i^{(\psi)} - X_i \beta\}] h_{\psi}(\psi; y_i) h_{,y}(\psi; y_i) \\ &\quad + h_{\psi; yy}(\psi; y_i) h_{,y}^{-1}(\psi; y_i) - h_{,yy}(\psi; y_i) h_{,y}^{-2}(\psi; y_i) h_{\psi; y}(\psi; y_i) \end{aligned} \quad (3.16)$$

The Jacobian $\varphi'_\theta = \partial\varphi/\partial\theta'$ is then given by

$$\varphi_{\theta'}(\theta) = \begin{bmatrix} \sum_1^n X_i' M_i(\theta; y_i) h_{;y}^{-1}(\hat{\psi}; y_i) \\ \sum_1^n M_i(\theta; y_i) h_{;y}^{-1}(\hat{\psi}; y_i) \{y_i^{(\hat{\psi})} - X_i \hat{\beta}\} \hat{\sigma}^{-1} \\ - \sum_1^n M_i(\theta; y_i) h_{;y}^{-1}(\hat{\psi}; y_i) h_\psi(\hat{\psi}; y_i) \end{bmatrix} \quad (3.17)$$

The overall information determinant recalculated on the φ scale is

$$|\hat{j}_{(\theta\theta)}| = |\hat{j}_{\theta\theta}||\varphi_{\{\theta\}}'(\hat{\theta})|^{-2} \quad (3.18)$$

and the nuisance information determinant recalculated on the φ scale is

$$|j_{(\lambda\lambda)}(\hat{\theta}_\psi)| = |j_{\lambda\lambda}(\hat{\theta}_\psi)||\varphi'_{\lambda'}(\hat{\theta}_\psi)\varphi_{\lambda'}(\hat{\theta}_\psi)|^{-1} \quad (3.19)$$

where we use the expressions (3.6) and (3.7).

For the scalar parameter $\chi(\theta)$ which measures departure from ψ in the φ scaling we use the row vector $\psi_{\varphi'}(\hat{\theta}_\psi)$ obtained as the final row vector in the inverse $\varphi_{\theta'}^{-1}(\hat{\theta}_\psi)$ of the Jacobian (3.17) evaluated at $\hat{\theta}_\psi$.

The third order significance function $p(\psi)$ for testing a ψ value is then obtained from (2.1) or (2.2) using the signed likelihood ratio r from (2.3) and the standardized maximum likelihood departure q from (2.4).

4. TWO EXAMPLES

Now consider two examples, one involving real data and one involving simulated data.

Example 1.

The survival times of animals in a 3×4 factorial experiment (Box & Cox, 1964) are recorded in Table 1; the first factor corresponds to three types of poison I, II, III and the second to four types of treatment A, B, C, D .

Table 1

Survival times of animals

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.25	0.31
	0.23	0.29	0.22	0.33

The model considered is an additive main effects model with normal error applied to the transformed variable $y^{(\psi)} = (y^\psi - 1)/\psi$. The data yields the maximum likelihood value $\hat{\psi} = -0.75$. The significance function $p(\psi)$ is recorded in Figure 1 together with the 95% confidence interval $(-1.165, -0.333)$; this is a standard confidence interval obtained from a pivotal quantity for ψ and its third order approximate distribution. The first order methods (Box & Cox, 1964) are obtained from the profile likelihood (1.3) and the Bayesian modifications of this. These are recorded in Figure 2; the likelihood 95% intervals is $(-1.13, -0.37)$ and the Bayesian interval is $(-1.18, -0.32)$. In this example the third order interval is larger and contains the first order intervals, thus correcting for apparent over precision in the likelihood methods.

The reciprocal transformation corresponding to $\psi = -1$ leads to a natural interpretation as the rate of dying. This value is within the confidence interval $(-1.165, -0.333)$ and the full analysis of variance using this value for ψ is recorded in Box & Cox (1964).

Example 2.

The power transformations y to y^ψ with $\psi \neq 0$ form a transformation group on the

positive real line. Thus, for testing a value $\psi = \psi_0$ with data y_1, \dots, y_n we can equivalently test a value $\psi = 1$ with transformed data $y_1^{\psi_0}, \dots, y_n^{\psi_0}$. Accordingly it suffices to examine the present inference procedure for the case where the true $\psi = 1$.

To assess the precision of inference for the parameter ψ in the context of the model $y^{(\psi)} = \mu + \sigma e$ with normal error, we consider a sample of 10 from the model with $\mu = 5$, $\sigma = 1$, $\psi = 1$ and standard normal errors; Due to the nature of the modified transformation (1.1) giving here $y^{(1)} = y - 1$, we have that the basic data variable y is normal $(6, 1)$. The data are

5.706884	5.738569	4.615973	4.305724	7.033090
5.797507	6.426300	7.285514	6.174499	5.536358

with maximum likelihood value $\psi = 1.322$. The third order 95% confidence interval is $(-2.355, 5.655)$; by comparison the first order confidence interval from $\ell_p(\psi)$ is $(-2.545, 5.425)$. With this small sample we note that the intervals are quite wide and that the third order procedure corrects a downward bias in the first order likelihood procedure from (2.3).

5. MONTE CARLO ASSESSMENT OF THE SIGNIFICANCE

The third order analysis in Section 3 produces a significance function $p(\psi)$ using either formula (2.1) or formula (2.2) calculated with the quantities r and q from (2.3) and (2.4). In repeated sampling from a model with $\psi = \psi_0$, the values of $p(\psi_0)$ should be approximately uniformly distributed and the values of $\Phi^{-1}\{p(\psi_0)\}$ should be approximately standard normal. To assess the accuracy of these approximations, we consider the simple linear model $y^{(\psi)} = \mu + \sigma e$ with standard normal error. We generate data with $n = 10$ and 20 , $\mu = 6$, $\sigma = 1$, $\psi = 1$ and calculate the significance $\Phi^{-1}\{p(1)\}$ using formulas (2.1) and (2.2). For $N = 10,000$ repetitions, the simulation results are given in Table 2. The simulated coverage probabilities using the third order procedure are closer to the nominal significance levels than those using the first order method when the sample size is small, i.e. $n = 10$. As the sample size is increased to $n = 20$, the difference between the two procedures is not as significant.

Table 2.

Estimated distribution function values calculated at familiar
nominal confidence values: using first order $p_1(\psi)$;
using 3rd order $p(\psi)$

	$n = 10$								
For $p(\psi)$	0.0260	0.0506	0.0987	0.1961	0.4967	0.7917	0.8899	0.9418	0.9667
For $p_1(\psi)$	0.0183	0.0388	0.0894	0.1936	0.5377	0.8539	0.9420	0.9743	0.9891
Nominal	0.025	0.05	0.1	0.2	0.5	0.8	0.9	0.95	0.975

	$n = 20$								
For $p(\psi)$	0.0278	0.0504	0.0996	0.1977	0.4959	0.7973	0.8986	0.9467	0.9711
For $p_1(\psi)$	0.0249	0.0465	0.0975	0.2031	0.5318	0.8417	0.9296	0.9691	0.9869
Nominal	0.025	0.05	0.1	0.2	0.5	0.8	0.9	0.95	0.975

We also present our simulation results as $q - q$ plots in Figure 3 and Figure 4 for the simple sizes $n = 10$ and 20 respectively. The horizontal scale records the result of the approximation using a normal scale and the vertical scale gives the simulation-based estimate of the exact value again expressed on a normal scale. Throughout most of its range the third order procedure stays close to the diagonal 45° line and closer for $n = 20$ than for $n = 10$. By contrast the likelihood ratio plot departs from zero at the center and has a systematic departure in slope from the diagonal 45° degree line; for values on the tails the nominal significance value over states the true significance and seems consistently non conservative.

For the third order an anomaly arises on the right tail, with actual significance being higher than the nominal. This anomaly is in a conservative direction and various procedures for correcting for it are being developed. This seems to correspond to data that have outliers on the left and are compressed on the right thus needing a large estimated ψ to balance the tails. A typical data array was extracted and after location scale standardization

gave

$$\begin{array}{ccccc} -2.610 & -0.717 & 0.060 & 0.316 & 0.329 \\ 0.377 & 0.408 & 0.522 & 0.564 & 0.750 \end{array}$$

After the $\hat{\psi} = 12.38$ power transformation of the original data the location scale standardized values are

$$\begin{array}{ccccc} -1.904 & -1.371 & -0.433 & 0.104 & 0.136 \\ 0.256 & 0.335 & 0.657 & 0.787 & 1.432 \end{array}$$

It would seem that a power transformation with a small sample does not provide a suitable rectification with a left tail outlier.

6. DISCUSSION

We have developed likelihood asymptotic methods for the Box & Cox (1964) problem and obtained a third order significance function for testing a scalar transformation parameter ψ in the model

$$y^{(\psi)} = X\beta + \sigma e$$

where the transformation $y_i^{(\psi)} = h(\psi; y_i)$ is coordinate by coordinate and is monotone increasing. As the Box & Cox model is far from having an exponential model parameterization we find as expected that the calculation of the maximum likelihood values $\hat{\theta}$ and $\hat{\theta}_\psi$ are the major computational barriers; but then these are also needed for the first order methods. These barriers for a particular data set are not a particular point of difficulty but for simulation studies they do require careful attention.

A rather extensive literature exists for the Box & Cox problem. In their paper (1964), Box and Cox suggested that after having chosen the transformation parameter, $\psi = \hat{\psi}$ say, one can make the estimation and inference under the usual normal theory on the chosen transformed scale.

Bickel & Doksum (1981) criticized the procedure, arguing that in some cases, for example when σ is small, the asymptotic variances of the estimates of the parameters β in the linear model are much larger when the transformation parameter ψ is unknown

than when it is a priori known, thus resulting in excess sensitivity in the approach. In response, Box and Cox argued that this variance inflation phenomenon is not relevant to their analysis, because they interpret model effects in terms of a known transformation, so that the parameter of interest is not the vector β related to the unknown ψ but rather a vector related to the given value $\hat{\psi}$. In order to remove the ψ -dependent scale effect, Box and Cox rescaled the response to obtain the working variables $z_i^{(\psi)}$. Hinkley and Runger(1984) gave a detailed analysis along these lines for the two data sets in Box and Cox's paper and found that the new model, using $z_i^{(\psi)}$ in place of $y_i^{(\psi)}$, was stable with respect to changes in $\hat{\psi}$.

Duan(1993) indicated that the z -transformation would not achieve the best possible reduction in the sensitivity if the design matrix was not symmetric as is the case with the textile data in Box and Cox (1964). He proposed a modification ζ of the z transformation as his working variable. If the distribution of X is symmetric, the analysis agrees with that of Box and Cox(1964).

Cox and Reid(1987) looked for a reparameterization of β and σ to make them orthogonal to the transformation parameter ψ , so that the maximum likelihood estimate of β does not depend strongly on ψ . They obtained results that differ slightly from those in the original Box & Cox paper.

The third order procedure has accuracy demonstrated by simulation; it thus removes the distributional uncertainty present with the first order likelihood or Bayesian analyses.

REFERENCES

- Barndorff-Nielsen, O.E. (1980). Conditionality resolutions. *Biometrika* **67**, 293–310.
- Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365.
- Barndorff-Nielsen, O.E. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557–563.
- Bickel, P.J. and Doksum, K.A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76**, 296–311.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *J. Royal Statist. Soc. B* **26**, 211–252.

Box, G.E.P. and Cox, D.R. (1982). An analysis of transformations revisited, rebutted. *J. of the Amer. Statist. Assoc.* **79**, 302–320.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Royal Statist. Soc. B* **49**, No. 1 1–39.

Duan, N. (1993) Sensitivity analysis for Box–Cox power transformation model : contrast parameters. *Biometrika* **80**, 4, 885–97.

Fraser, D.A.S. (1964). Local conditional sufficiency. *J. Royal Statist. Soc. B* **26**, 52–62.

Fraser, D.A.S. (1979). *Inference and Linear Models*. New York: McGraw Hill.

Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77**, 333–341.

Fraser, D.A.S. and Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica* **47**, 33–53.

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

Lugannani, R. and Rice, S.O (1980). Saddlepoint approximation for the distribution of the sums of independent random variables. *Adv. Appl. Prob.* **12**, 475–490.

Figure 1. Third order significance function for the Box & Cox data on survival time, together with the 95% confidence interval.

Figure 2. First order profile likelihoods for the Box & Cox data on survival time together with the corresponding 95% ranges.

Figure 3. With $N = 10,000$, simulation of the $z = \Phi^{-1}(p)$ values obtained from the first order and the third order significance procedures: $q - q$ plot with $n = 10$.

Figure 4. With $N = 10,000$, simulation of the $z = \Phi^{-1}(p)$ values obtained from the first order and the third order significance procedures: $q - q$ plot with $n = 20$.