POWER TRANSFORMATIONS WHEN FITTING
THEORETICAL MODELS TO DATA

Raymond J. Carroll
and
David Ruppert

Running Title: Fitting Theoretical Models

1 National Heart, Lung and Blood Institute and the University of North
Carolina. Supported by the Air Force Office of Scientific Research
Grant AFOSR-80-0080.

2 University of North Carolina. Supported by National Science Foundation
Grant MCS 8100748.

Some key words: Transformations, Box-Cox models, theoretical models,
robustness.

AMS 1970 Subject Classifications: Primary 62F20; Secondary 62G35.
We investigate power transformations in non-linear regression problems when there is a physical model for the response but little understanding of the underlying error structure. In such circumstances and unlike the ordinary power transformation model, both the response and the model must be transformed simultaneously and in the same way. We show by an asymptotic theory and a small Monte-Carlo study that for estimating the model parameters there is little cost for not knowing the correct transform a priori; this is in dramatic contrast to the results for the usual case that only the response is transformed.
1: INTRODUCTION

Often in scientific work, one observes data \( y \) and \( x^i = (x^i_1 \ldots x^i_p) \) and postulates that these data follow a model

\[
y^i = f(x^i, \theta_0), \quad i = 1, \ldots, N,
\]

where \( \theta_0 \) is a \( k \)-parameter vector. The function \( f \) may be derived, for example, from differential equations believed to govern the physical system which gave rise to the data. The deterministic model (1.1) is often inadequate since the data exhibit random variation, but whereas \( f \) was derived from theoretical considerations, there is really no firm understanding of the mechanism producing the randomness. In this case, one typically assumes that

\[
y^i = f(x^i, \theta_0) + \epsilon^i,
\]

where the \( \{\epsilon^i\} \) are i.i.d. \( N(0, \sigma^2_0) \). In those cases in which the data suggest that model (1.2) is also unsatisfactory, one might then assume that the errors are multiplicative and log-normal, so that

\[
\log(y^i) = \log(f(x^i, \theta_0)) + \epsilon^i.
\]

The point here is that model (1.1) is equivalent to the model

\[
h(y^i) = h(f(x^i, \theta_0))
\]

whenever \( h(\cdot) \) is a monotonic transformation. Therefore (1.2) and (1.3) are based on the same theoretical model, but they allow variability into the model in different fashions.
A more flexible approach is to take a sufficiently rich family of strictly monotonic transformations \( h(y, \lambda) \), indexed by the \( m \)-vector parameter \( \lambda \), and to assume that for some value \( \lambda_o \).

\[
(1.4) \quad h(y_i, \lambda_o) = h(f(x_i, \theta_o), \lambda_o) + \epsilon_i.
\]

The model (1.4) is in the spirit of Box and Cox (1964), who suggested the family of power transformations with \( m = 1 \) and

\[
(1.4b) \quad h(y, \lambda) = y(\lambda) = (y^\lambda - 1)/\lambda \quad \text{if } \lambda \neq 0 \\
= \log(y) \quad \text{if } \lambda = 0.
\]

However, as we will make clear, our proposed model (1.4) has greatly different ramifications than usually associated with the power family. Box and Cox (1964) used their family in a study of the transformation model

\[
(1.5) \quad h(y, \lambda_o) = x^T \theta_o + \epsilon.
\]

Notice here that, unlike (1.4), the regression function in (1.5) is not transformed. Box and Cox sought a transformation which achieves 1) a simple, additive or linear model, 2) homoscedastic errors and 3) normally distributed errors. Our model is different. Theoretical considerations already provide a regression function. We hope to transform the response \textit{and} the regression function simultaneously to obtain homoscedasticity and normality.

There are two reasons for using model (1.4) instead of simply fitting (1.1) by least squares or some other method. First, estimation of \( \theta_o \) based on model (1.4) should be more efficient than other methods. Second, it may
be necessary to estimate the entire conditional distribution of \( y \) given \( x \); if the data clearly suggest that the distributions of \( \{y_i - f(x_i, \theta_0)\} \) are not constant, one must go beyond standard regression methodology.

An example, which partly motivated the research of this paper, concerns the relationship between egg production in a fish stock and subsequent recruitment into the stock. At least for some species, as egg production increases, the change in the skewness and variance of recruitment is as large as the change in the median recruitment, and this change in distributional shape may have important implications for management of the fishery.

The outline of the paper is as follows. Section 2 discusses a current controversy concerning the model of Box and Cox. Bickel and Doksum (1981) have shown that, in model (1.5), the ML estimate of \( \theta_0 \) can be much more variable when \( \lambda_0 \) is estimated compared to when \( \lambda_0 \) is known. In Section 3, we demonstrate for our model (1.5) an entirely different result: the ML estimate of \( \theta_0 \) in model (1.4) turns out to be only slightly more variable when \( \lambda_0 \) is unknown compared to when \( \lambda_0 \) is known. In Section 4 we prove a considerably stronger result. By examining a weighted least absolute deviations estimator, we provide a lower bound of \( 2/\pi \) on the asymptotic relative efficiency of the ML estimator of \( \theta_0 \) in model (1.4) when \( \lambda_0 \) is unknown compared to the MLE when \( \lambda_0 \) is known.

2: RECENT STUDIES OF THE BOX AND COX MODEL

In Section 7 of Box and Cox's original paper they discuss the analysis of effects after transformation. They state that, after finding \( \hat{\lambda} \), one should estimate effects (regression parameters) on the scale \( \hat{\lambda} \) which has been chosen for analysis and not on the true but unknown \( \lambda_0 \) scale. However, in discussing interactions, they go on to state that "The general conclusion will be that to allow for the effect of analysing in terms of \( \hat{\lambda} \) rather than \( \lambda_0 \), the
residual degrees of freedom need only be reduced by ... the number of component parameters in \( \lambda \).

Box and Tiao (1968) agree, stating that the only practical effect between using \( \hat{\lambda} \) in the posterior distribution of \( \theta_0 \), rather than the true \( \lambda_0 \), is an adjustment in the degrees of freedom.

Bickel and Doksum (1981) disagree with this conclusion. Following calculations for the location problem done by Hinkley (1975) and suggestive Monte-Carlo results of Spitzer (1978) and Carroll (1980), they calculated for general regression the large sample information matrix of \( \lambda_0, \sigma_0^2 \) and \( \theta_0 \). They found that the large sample variance of \( \hat{\theta} \) is larger, often much larger, when \( \lambda_0 \) is estimated compared to when \( \lambda_0 \) is known. They also state that the conclusion of Box and Tiao is not correct. On a technical level, part of the discrepancy between Bickel and Doksum's and Box and Tiao's results may be due to the use of different transformations. Bickel and Doksum use (1.4b), while Box and Tiao use

\[
z(\lambda) = y(\lambda)/\dot{y}^{\lambda-1},
\]

where \( \dot{y} \) is the geometric mean of the \( \{y_i\} \). However, Hinkley and Runger (1982) found \( z(\lambda) \) unsatisfactory in several respects. The differences may also be contextual; at the null hypothesis of no interaction effects, one \textit{can} act as if \( \lambda_0 \) were known, with an appropriate change in the degrees of freedom. See Carroll (1982) and Doksum and Wong (1981).

Since power transformations have been used often and with real satisfaction by applied statisticians, the findings of Bickel and Doksum were surprising and led to further research. Hinkley and Runger argue that the parameter \( \theta_0 \) in (1.5) is not physically meaningful; it is defined in an unknown scale \( \lambda_0 \) so that a unit change in \( x \) is not easily interpreted by \( \theta_0 \) alone. Instead, they argue that in practice, the relevant distribution is the conditional distribution of \( \hat{\theta} \) given \( \hat{\lambda} \). As \( N \to \infty \), the conditional variance of \( \hat{\theta} \) given \( \hat{\lambda} \)
and the variance of $\hat{\theta}$ when $\lambda_o$ is known converge to the same matrix. They then argue that, when analyzing $\theta_o$, no adjustment need be made for the fact that $\lambda_o$ was estimated. This appealing behavior is somewhat counter-balanced by difficulties with the conditional mean in hypothesis testing in unbalanced designs, as pointed out by Carroll (1982).

Carroll and Ruppert (1981) also noticed the difficulty with interpreting $\theta_o$ and studied predicting the median of $y$ on the original data scale by backtransforming $\hat{x}^T \hat{\theta}$. This idea of looking at the response surface avoids the problems of definition inherent with $\theta_o$ being defined in an unknown or data dependent scale. They found that when predicting the median of $y$, the effect of not knowing $\lambda_o$ can be large but is in general similar to the effect of adding one more regression parameter, and it is certainly much less severe than the effect when estimating $\theta_o$.

The above discussion establishes the extent of the controversy surrounding the Box and Cox model applied (1.5). We believe (1.4) entirely avoids this controversy. First, the parameter $\theta_o$ has physical meaning even if $\lambda_o$ is unknown, since $f(x_i^T \theta_o)$ is the median of $y_i$ no matter what the true scale. Secondly, the large sample analysis to follow indicates that $\hat{\theta}$ is only slightly more variable when $\lambda_o$ is estimated than when $\lambda_o$ is known.

3: LIKELIHOOD ANALYSIS

The likelihood analysis proceeds as follows: define

$$
\psi_i = \frac{\partial h(\psi_i, \theta_o, \lambda_o)}{\partial \theta_o}
$$

$$
f_{\psi_i}(\theta) = f(x_i^T \theta), \quad f_{\psi_i} = f_{\psi_i}(\theta_o),
$$

$$
h_{y}(y) = h_y(y, \lambda) = \frac{\partial h(y, \lambda)}{\partial y}, \quad \text{and} \quad h_{y}(y) = h(y, \lambda).
$$

Let $h_{\lambda}(y)$ and $h_{\lambda\lambda}(y)$ be the gradient vector and Hessian of $h(y, \lambda)$ with respect to $\lambda$. By simple algebra we find the joint information matrix of $(\theta_o, \sigma_o, \lambda_o)$
as (all summations are from 1 to N)

\[ N^{-1} I = \begin{bmatrix} S/\sigma_0^2 & 0 & C_1/\sigma_0^2 \\ \vdots & 1/(2\sigma_0^2) & C_2/\sigma_0^2 \\ \vdots & \vdots & C_3/\sigma_0^2 \end{bmatrix} \]

where

\begin{align*}
S &= N^{-1} \sum \epsilon_i z_i \epsilon_i^t \\
C_1 &= -N^{-1} E[\sum \epsilon_i [h_\lambda(y_i) - h_\lambda(\tilde{f}_i)]^t] \\
C_2 &= -N^{-1} E[\sum \epsilon_i [h_\lambda(y_i) - h_\lambda(\tilde{f}_i)]^t] \\
C_3 &= N^{-1} E\{[h_\lambda(y_i) - h_\lambda(\tilde{f}_i)] [h_\lambda(y_i) - h_\lambda(\tilde{f}_i)]^t \\
&\quad + \epsilon_i [h_\lambda(y_i) - h_\lambda(\tilde{f}_i)] + (\partial/\partial \lambda)(\partial/\partial \lambda)^t \log[h_y(y_i)]\}.
\end{align*}

Using the work of Hoadley (1971), it is straightforward, though perhaps somewhat tedious, to establish conditions sufficient that \((\hat{\theta}, \hat{\sigma}^2, \hat{\lambda})\) is consistent and asymptotically normal. We will not pursue this matter further, but rather we will assume that \((\hat{\theta}, \hat{\sigma}^2, \hat{\lambda})\) is approximately \(N((\theta_0, \sigma_0^2, \lambda_0)^t, I^{-1})\) and we will study \(I^{-1}\).

In general, \(C_1\) and \(C_2\) are not zero and the asymptotic distribution of \((\hat{\lambda}, \hat{\sigma}^2)\) when \(\lambda_0\) is estimated differs from when \(\lambda_0\) is known. At least to this point then, the analysis is similar to those done in the usual Box-Cox model (1.5). The key question, of course, is whether or not \(C_1\) and \(C_2\) are sufficiently different from zero to seriously affect the distribution of \(\hat{\lambda}\).

The expressions \(C_1, C_2\) and \(C_3\) are complex even when \(f_\varepsilon(\theta_0)\) has a nice form such as simple linear regression. To simplify matters sufficiently that we can gain some insight about the difference between knowing and estima-
ting $\lambda_0$, we follow Bickel and Doksum and others and let $\sigma_0 + 0$. While Bickel and Doksum let $N \to \infty$ and $\sigma_0 \to 0$ simultaneously, we let $N \to \infty$ and then $\sigma_0 \to 0$. There is no essential difference between the two approaches. Our is very suitable for heuristic arguments.

It should be emphasized that we are not concerned only, or even primarily, with small $\sigma_0$. In fact, the need for transformation is greater when $\sigma_0$ is large. The small $\sigma_0$ asymptotics do, however, lead to major simplifications, and the Monte-Carlo results presented later agree with them.

Taylor expansions show that under mild regularity conditions

\begin{equation}
(3.3) \quad \begin{align*}
C_1 &= o(\sigma^2_0), \\
C_2 &= o(\sigma^2_0), \\
C_3 &= o(\sigma^2_0) \text{ as } \sigma_0 \to 0.
\end{align*}
\end{equation}

Standard calculations show that when $\lambda_0$ is known,

\begin{equation}
(3.4) \quad N^{-k} \text{ Covariance } [(\hat{\theta} - \theta_0)/\sigma_0, (\hat{\sigma}^2 - \sigma^2_0)/\sigma_0^2|\lambda_0 \text{ known}] \to A^{-1} = \begin{bmatrix} S^{-1} & 0 \\ 0 & 2 \end{bmatrix}.
\end{equation}

Let $D = \text{Diag}(\sigma_0, \sigma^2_0, 1)$. Then, to find this limiting covariance matrix when $\lambda_0$ is unknown, we must find the upper left $(k + 1) \times (k + 1)$ corner of

\begin{equation}
DID = \begin{bmatrix}
S & 0 & C_1/\sigma_0 \\
\cdot & \frac{1}{2} & C_2/\sigma^2_0 \\
\cdot & \cdot & C_3/\sigma^2_0
\end{bmatrix}
\end{equation}

which by standard results on inverting partitioned matrices is

\begin{equation}
A^{-1} + FE^{-1}F^t
\end{equation}

where $A^{-1}$ is given in (3.4),

\begin{equation}
E = C_3/\sigma^2_0 - B^t A B,
\end{equation}

\begin{equation}
F = A^{-1} B,
\end{equation}

\begin{equation}
D = \text{Diag}(\sigma_0, \sigma^2_0, 1).
\end{equation}
and
\[ B = \begin{pmatrix} C_1/\sigma_0 \\ C_2/\sigma_0^2 \end{pmatrix} \]
Clearly,
\[ F = \begin{pmatrix} S^{-1} C_1/\sigma_0 \\ 2C_2/\sigma_0 \end{pmatrix} \]
and
\[ E = C_2/\sigma_0^2 - C_1^t S^{-1} C_1/\sigma_0^2 - 2C_2^t C_2/\sigma_0^4. \]

In order to obtain simple asymptotics, we will assume that for \( \sigma_0 \) fixed, 
\( C_1/\sigma_0, C_2/\sigma_0^2 \), and \( C_3/\sigma_0^2 \) converge as \( N \to \infty \), and that these, in turn, have limits \( D_1, D_2, \) and \( D_3 \) respectively as \( \sigma_0 \to 0 \). We also assume that 
\( S \to S_\infty \) (positive definite) as \( N \to \infty \). If \( D_3 - 2D_2^t D_2 \) is nonsingular, then
\[
\lim_{\sigma_0 \to 0} \lim_{N \to \infty} F E^{-1} p t = \begin{pmatrix} 0 & 0 \\ 0 & W \end{pmatrix}
\]
where \( W = 4 D_2^t \left[ D_3 - D_2^t D_2 \right]^{-1} D_2 \). Therefore
\[
\lim_{\sigma_0 \to 0} \lim_{N \to \infty} \left( A^{-1} + F E^{-1} F^t \right) = \begin{pmatrix} S^{-1} & 0 \\ 0 & 2 + W \end{pmatrix}
\]

THEOREM 1. Assume that the limits \( D_1, D_2, D_3, S_\infty \) mentioned above exist and that \( D_3 - 2D_2^t D_2 \) is nonsingular. As \( N \to \infty \) and then \( \sigma_0 \to 0 \), the limit distribution of \( \hat{\theta} \) is the same whether \( \lambda_0 \) is known or unknown. The limit distribution of \( \hat{\sigma} \) depends on whether \( \lambda_0 \) is known or unknown.

As an example consider multiple linear regression and the power transformation family, i.e., \( h(y, \lambda) \) is given by (1.4b) and
\[ h(y, \lambda) = x_t^t \theta_0 + \epsilon \]
where \( x_1, \ldots, x_n \) are known \( k \times 1 \) vectors. Also, suppose that \( \lambda_0 = 0 \), i.e., the log transformation is needed. Then \( h_y(y) = y^{\lambda_0 - 1} \), \( h_\lambda(y) = (\log y)^{3/2} \), and \( h_{\lambda\lambda}(y) = (\log y)^3/3 \). We find that

\[
A = N^{-1} \sum x_i x_i^T/(x_i^T \theta_0)^2
\]

\[
C_1 = -(2N)^{-1} \sum \{\log(x_i^T \theta_0) + \epsilon_i\}^2 - (\log(x_i^T \theta_0))^2
\]

\[
C_2 = -(2N)^{-1} \sum \epsilon_i^2 \{\log(x_i^T \theta_0) + \epsilon_i\}^2 - (\log(x_i^T \theta_0))^2
\]

\[
C_3 = (4N)^{-1} \sum \{\log(x_i^T \theta_0) + \epsilon_i\}^3 - (\log(x_i^T \theta_0))^3
\]

\[
= \gamma/4 \sigma_0^4 + 2 \sigma_0^2/N \sum (\log(x_i^T \theta_0))^2
\]

Therefore,

\[
D_1 = \lim_{N \to \infty} (2N)^{-1} \sum x_i/(x_i^T \theta_0)
\]

\[
D_2 = \lim_{N \to \infty} N^{-1} \sum \log(x_i^T \theta_0)
\]

and

\[
D_3 = 2 \lim_{N \to \infty} N^{-1} \sum (\log(x_i^T \theta_0))^2
\]

provided the above limits exists. Thus, the \( 1 \times 1 \) matrix \( D_3 - 2D_2^T D_2 \) is twice the limit of the variance of \( \log(x_1^T \theta_0), \ldots, \log(x_N^T \theta_0) \), and will be nonsingular except in degenerate situations.

There is thus a fundamental difference between the models (1.4) and (1.5). A small simulation study is outlined in Section 6 and helps back up Theorem 1. This result can be extended to non-normal error distributions as
well as the robust methods of Carroll (1980) and Bickel and Doksum (1981). The details are not instructive.

4: A LOWER BOUND ON THE EFFICIENCY OF THE MLE.

Let $\hat{\theta}(\hat{\lambda})$ and $\hat{\theta}(\lambda_o)$ denote the ML estimator with $\lambda_o$ estimated and known respectively. Let $\text{ARE}(\hat{\theta}_1, \hat{\theta}_2)$ be the asymptotic relative efficiency of $\hat{\theta}_1$ to $\hat{\theta}_2$. For fixed $\sigma_o$, it is difficult to find $\text{ARE}(\hat{\theta}(\lambda_o), \hat{\theta}(\hat{\lambda}))$ and, in fact, this may depend on $\theta_o$, $\lambda_o$, the $\{x_i\}$ and the coordinate of $\theta$ being estimated. All that can be said for certain is that this ARE is at least one and converges to one as $\sigma_o \to 0$. In this section we will define a weighted $L_1$ or least absolute deviation estimator $\hat{\theta}(\nu)$ and show that $\text{ARE}(\hat{\theta}(\lambda_o), \hat{\theta}(\nu)) \leq \pi/2$. Under reasonable regularity conditions, this means that $\text{ARE}(\hat{\theta}(\lambda_o), \hat{\theta}(\hat{\lambda}))$ is bounded between one and $\pi/2$, in vivid contrast to the Box and Cox model (1.5) in which this last ARE can approach infinity. We first look at general weighted $L_1$ estimators. The results stated here seem to be new and are of interest in their own right.

Let $\omega_1, \ldots, \omega_N$ be positive numbers and let $\theta(L)$ be any point which minimizes the expression

$$\sum \omega_i |y_i - f_i(\hat{\theta}(L))| .$$

Under (1.4), $f_i(\theta_o)$ is the unique median of $y_i$, so we can expect $\hat{\theta}(L)$ to be consistent. The unweighted $L_1$ estimate for linear models was studied by Ruppert and Carroll (1980). Those results suggest that
Define \( r_i = y_i - f_i(\theta_o) \) and let \( m_i \) be the density of \( r_i \). By a generalization of the strong law, for example Theorem 7.1 of Carroll and Ruppert (1982) which itself generalizes Lemma 4.2 of Bickel (1975),

\[
\begin{align*}
0 &= \sum_i \omega_i \{ \text{sign}(y_i - f_i(\hat{\theta}(L))) - \text{sign}(r_i) \} s_i \\
&\quad - \mathbb{E}[\sum_i \omega_i \{ \text{sign}(y_i - f_i(\theta)) - \text{sign}(r_i) \} s_i ] | \theta = \hat{\theta}(L)).
\end{align*}
\]

Now, as \( \epsilon \to 0 \), we obtain that

\[
\mathbb{E}(\text{sign}(r_i + \epsilon) - \text{sign}(r_i)) \to 0.
\]

Combining (4.1)-(4.3) we get to order \( o(N^{-\frac{1}{2}}) \),

\[
(\hat{\theta}(L) - \theta_o) \approx \frac{1}{\sum_t \omega_t m_t(o) s_t s_t^t} \sum_t \omega_t s_t \text{sign}(r_t).
\]

Now, since for model (1.4)

\[
\epsilon \approx h(f_i(\theta_o) + r_i, \lambda_o) - h(f_i(\theta_o), \lambda_o),
\]

we then have

\[
m_i(0) = (2\pi \sigma_o^2)^{-\frac{1}{2}} h_y(f_i(\theta_o), \lambda_o).
\]
Thus, if we chose

\begin{equation}
\omega \_z = h \_y (f \_z (0 \_o), \lambda \_o),
\end{equation}

we have by (4.4)-(4.6) and the Central Limit Theorem that

\[
N^\frac{1}{2} (\hat{\theta}(L) - \theta \_o) / \sigma \_o \xrightarrow{L} N(0, (\pi/2) S^{-1}).
\]

Now \( \hat{\theta}(L) \) is not a bona fide estimator since \( \omega \_z \) in (4.6) requires \( \lambda \_o, \theta \_o \) to be known. However, if in (4.6) one plugs in any \( N^\frac{1}{2} \) consistent estimators of \( \theta \_o \) and \( \lambda \_o \) and calls the \( L \) estimate based on these new weights \( \hat{\theta}(\omega) \), then using Theorem 7.1 of Carroll and Ruppert (1982), one can also show that

\[
N^\frac{1}{2} (\hat{\theta}(\omega) - \theta \_o) / \sigma \_o \xrightarrow{L} N(0, (\pi/2) S^{-1}).
\]

Now, because

\[
N^\frac{1}{2} (\hat{\lambda} \_o - \lambda \_o) / \sigma \_o \xrightarrow{L} N(0, S^{-1}).
\]

it then follows that

\[
\text{ARE}(\hat{\theta}(\lambda \_o), \hat{\theta}(\omega)) = \pi/2,
\]

\begin{equation}
\text{ARE}(\hat{\theta}(\lambda \_o), \hat{\theta}(\lambda)) \leq \pi/2.
\end{equation}

Theorem 1 and the Monte-Carlo results to follow indicate that the upper bound in (4.8) is quite conservative. The beauty of (4.8) is that it is a bound that does not depend on \( \sigma \_o \).

The weighted \( L \) estimator may well be useful for example if in (1.4) one suspected that the errors \( \{\epsilon \_z\} \) are not normal. It is a consistent estimator of \( \theta \_o \) provided that 0 is the unique median of \( \epsilon \_z \). Symmetry of \( \epsilon \_z \) is not needed.
5: THE K-SAMPLE PROBLEM

Our model (1.4) and Theorem 1 provide some useful insight into the k-sample problem under the formulation (1.5) of Box and Cox. In their model, for each of k populations we have

\[ h(y_{i,j}, \lambda) = \mu_j + \epsilon_{i,j} \quad j = 1, \ldots, k; \quad i = 1, \ldots, N_i. \]

The equivalent formulation from our viewpoint is

\[ h(y_{i,j}, \lambda) = h(\xi_j, \lambda) + \epsilon_{i,j}, \]
\[ \mu_j = h(\xi_j, \lambda). \]

Here \( \xi_j \) is the median of \( y_{i,j} \) on the original scale and \( \mu_j \) is the expected value of \( y_{i,j} \) in the \( \lambda \) scale. The results of Carroll and Ruppert imply that for estimating the \( \xi \)'s, there is little cost in not knowing \( \lambda \), while for estimating the \( \mu \)'s, Bickel and Doksum show that the cost of not knowing \( \lambda \) can be enormous. Since

\[ H_0 : \mu_1 = \ldots = \mu_k \quad \text{iff} \quad H_0 : \xi_1 = \ldots = \xi_k, \]

there should be little cost in testing for equality of means when \( \lambda \) is unknown. These heuristics are formally proven by Carroll (1982) and Doksum and Wong (1981).

6: MONTE-CARLO.

To study \( \hat{\theta} \) when \( N \) is finite and \( \sigma \) is not necessarily small, we undertook a small simulation of the model

\[ h(y_{i}, \lambda) = h(\theta_1 + \theta_2 x_{i}, \lambda) + \sigma \epsilon_{i}, \]
where \( h() \) is the Box and Cox power family (1.4b). In our simulations, \( N = 50 \), the design points \( \{x_i\} \) were equally spaced on \([-1, 1]\), the errors were normally distributed with mean zero and variance one and \( \theta_1 = 7, \theta_2 = 2 \). We considered three estimators:

1) ML estimator, \( \lambda_o \) known (KNOWN)
2) ML estimator, \( \lambda_o \) unknown (MLE)
3) The ordinary least squares estimator (LSE) without any transformation.

The median of \( y \) is \( \theta_1 + \theta_2 x \), so that LSE forms an especially plausible estimator of the slope \( \theta_2 \) (for which it is consistent). We chose three values of \( \sigma_o \):

\[ \sigma_o = 0.05, 0.10, \text{and} 0.50. \]

We present results in Tables 1 and 2 for \( \lambda_o = 0 \) (log-normal data) and \( \lambda_o = 0.25 \). There were 600 replications of the experiment for each \( (\lambda_o, \sigma_o) \) and each estimator, all generated from a common set of random numbers. The normal random deviates were generated from the IMSL routine GGNPM. Estimation of \( (\theta_1, \theta_2) \) for each \( \lambda \) was done by the IMSL routine ZXSSQ while ZXGSN was used to estimate \( \lambda_o \).

The results for the ML estimator with \( \lambda_o \) unknown (denoted MLE) are very encouraging. The mean square errors for MLE are quite close to those for KNOWN, the ML estimator with \( \lambda_o \) known, especially for the slope \( \theta_2 \). These results agree with our small \( \sigma \) theory and indicate the minimal cost for not knowing \( \lambda_o \). The relative efficiencies of MLE to KNOWN are always well above the lower bound of \( 2/\pi \). To appreciate how well MLE does relative to KNOWN
(line 2 of Tables 1 and 2), it is enlightening to study Table 5 of Bickel and Doksum (1981); in their model which we call (1.5), they have ratios MLE($\lambda_0$ estimated)/KNOWN($\lambda_0$ known) always at least 1.5 and as large as 211, while ours never exceed 1.2.

The other valuable point learned from Table 2 is that when estimating the slope $\theta_2$, the ML estimator MLE with $\lambda_0$ unknown tends to dominate the LSE, especially for larger values of $\sigma_0$. In other words, for our model (1.4), there is real value to transformation when it is appropriate.
REFERENCES


TABLE #1

Results of the Monte-Carlo study described in the text. These results are for the INTERCEPT. The median response is linear with intercept = 7 and slope = 2.

**KNOWN** = ML estimate with $\lambda$ known.
**MLE** = ML estimate with $\lambda$ unknown.
**LSE** = ordinary least squares estimate.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>0.00</th>
<th>0.10</th>
<th>0.50</th>
<th>0.05</th>
<th>0.10</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$</td>
<td>0.05</td>
<td>0.10</td>
<td>0.50</td>
<td>0.05</td>
<td>0.10</td>
<td>0.50</td>
</tr>
<tr>
<td>BIAS OF KNOWN</td>
<td>0.03</td>
<td>0.06</td>
<td>0.56</td>
<td>0.01</td>
<td>0.03</td>
<td>0.23</td>
</tr>
<tr>
<td>MSE OF KNOWN</td>
<td>2.41</td>
<td>9.67</td>
<td>24.87</td>
<td>0.90</td>
<td>3.59</td>
<td>9.04</td>
</tr>
<tr>
<td>BIAS OF MLE</td>
<td>0.02</td>
<td>0.04</td>
<td>0.60</td>
<td>0.01</td>
<td>0.02</td>
<td>0.19</td>
</tr>
<tr>
<td>MSE OF MLE/MSE OF KNOWN</td>
<td>1.02</td>
<td>1.05</td>
<td>1.14</td>
<td>1.01</td>
<td>1.03</td>
<td>1.12</td>
</tr>
<tr>
<td>MSE OF MLE - MSE OF KNOWN</td>
<td>0.05</td>
<td>0.47</td>
<td>3.44</td>
<td>0.01</td>
<td>0.09</td>
<td>1.09</td>
</tr>
<tr>
<td>S.E. OF ABOVE DIFF.</td>
<td>0.02</td>
<td>0.15</td>
<td>0.77</td>
<td>0.01</td>
<td>0.04</td>
<td>0.25</td>
</tr>
<tr>
<td>BIAS OF LSE</td>
<td>0.11</td>
<td>0.40</td>
<td>9.48</td>
<td>0.04</td>
<td>0.13</td>
<td>2.60</td>
</tr>
<tr>
<td>MSE OF MLE/MSE OF LSE</td>
<td>0.97</td>
<td>0.90</td>
<td>0.22</td>
<td>1.00</td>
<td>0.98</td>
<td>0.63</td>
</tr>
<tr>
<td>MSE OF MLE - MSE OF LSE</td>
<td>-0.06</td>
<td>-1.15</td>
<td>-96.62</td>
<td>0.00</td>
<td>-0.06</td>
<td>-6.07</td>
</tr>
<tr>
<td>S.E. OF ABOVE DIFF.</td>
<td>0.04</td>
<td>0.33</td>
<td>4.71</td>
<td>0.01</td>
<td>0.06</td>
<td>0.78</td>
</tr>
</tbody>
</table>

In these calculations, the mean square error (MSE) and S.E. of difference terms are multiplied by $T^2$. Here $T = 10$ if $\sigma \leq 0.10$, $T = 1$ if $\sigma = 0.50$. 
TABLE #2

Results of the Monte-Carlo study described in the text. These results are for the SLOPE. The median response is linear with intercept = 7 and slope = 2.

- **KNOWN** = ML estimate with $\lambda$ known.
- **MLE** = ML estimate with $\lambda$ unknown.
- **LSE** = ordinary least squares estimate.

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>0.00</th>
<th>0.50</th>
<th>0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$</td>
<td>0.05</td>
<td>0.10</td>
<td>0.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0.05</th>
<th>0.10</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS OF KNOWN</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>MSE OF KNOWN</td>
<td>7.08</td>
<td>28.36</td>
<td>72.23</td>
</tr>
<tr>
<td>BIAS OF MLE</td>
<td>-0.01</td>
<td>-0.04</td>
<td>-0.15</td>
</tr>
<tr>
<td>MSE OF MLE</td>
<td>1.06</td>
<td>1.06</td>
<td>1.01</td>
</tr>
<tr>
<td>MSE OF KNOWN</td>
<td>1.06</td>
<td>1.06</td>
<td>1.03</td>
</tr>
<tr>
<td>MSE OF MLE - MSE OF KNOWN</td>
<td>0.41</td>
<td>1.57</td>
<td>0.95</td>
</tr>
<tr>
<td>S.E. OF DIFF.</td>
<td>0.10</td>
<td>0.40</td>
<td>0.67</td>
</tr>
<tr>
<td>BIAS OF LSE</td>
<td>0.05</td>
<td>0.15</td>
<td>2.97</td>
</tr>
<tr>
<td>MSE OF MLE</td>
<td>0.98</td>
<td>0.96</td>
<td>0.59</td>
</tr>
<tr>
<td>MSE OF LSE</td>
<td>1.01</td>
<td>1.01</td>
<td>0.91</td>
</tr>
<tr>
<td>MSE OF MLE - MSE OF LSE</td>
<td>-0.16</td>
<td>-1.29</td>
<td>-50.54</td>
</tr>
<tr>
<td>S.E. OF DIFF.</td>
<td>0.18</td>
<td>0.80</td>
<td>5.10</td>
</tr>
</tbody>
</table>

In these calculations, the mean square errors (MSE) and S.E. of difference terms are multiplied by $T^2$. Here, $T = 10$ if $\sigma \leq 0.10$, $T = 1$ if $\sigma = 0.50$. 