POWER TRANSFORMATIONS WHEN FITTING THEORETICAL MODELS TO DATA

Raymond J. Carroll[1]

and

David Ruppert[2]

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.

ABSTRACT

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 $x^t = (x_1 \, . \, . \, x_p)$ and postulates that these data follow a model

$$(1.1) \qquad y_i = f(x_i, \theta_o), \; i = 1, \ldots, N,$$

where θ_o 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

$$(1.2) \qquad y_i = f(x_i, \theta_o) + \epsilon_i \, ,$$

where the $\{\epsilon_i\}$ are i.i.d. $N(0,\sigma_o^2)$. 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

$$(1.3) \qquad log(y_i) = log(f(x_i, \theta_o)) + \epsilon_i \, .$$

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

$$h(y_i) = h(f(x_i, \theta_o))$$

whenever $h(\bullet)$ 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 λ, and to assume that for some value $\lambda_\circ$.

$$(1.4) \qquad h(y_i,\ \lambda_\circ) = h(f(x_i,\ \theta_\circ),\ \lambda_o) + \epsilon_i.$$

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

$$(1.4b) \qquad \begin{aligned} h(y,\lambda) = y^{(\lambda)} &= (y^{\lambda} - 1)/\lambda \quad \text{if } \lambda \neq 0 \\ &= log\ (y) \quad \text{if } \lambda = 0. \end{aligned}$$

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

$$(1.5) \qquad h(y,\lambda_o) = x^t\ \theta_o + \epsilon.$$

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 *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 θ_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 $\{y_i - f(x_i, \theta_o)\}$ 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 θ_o can be much more variable when λ_o is estimated compared to when λ_o is known. In Section 3, we demonstrate for our model (1.5) an entirely different result: the ML estimate of θ_o in model (1.4) turns out to be only slightly more variable when λ_o is unknown compared to when λ_o 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 θ_o in model (1.4) when λ_o is unknown compared to the MLE when λ_o 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 λ_o 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 λ_o, the residual degrees of freedom need only be reduced by ... the number of component parameters in λ".

Box and Tiao (1968) agree, stating that the only practical effect between using $\hat{\lambda}$ in the posterior distribution of θ_o, rather than the true λ_o, 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 λ_o, σ_o^2 and θ_o. They found that the large sample variance of $\hat{\theta}$ is larger, often much larger, when λ_o is estimated compared to when λ_o 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 *can* act as if λ_o 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 θ_o in (1.5) is not physically meaningful; it is defined in an unknown scale λ_o so that a unit change in x is not easily interpreted by θ_o 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 λ_o is known converge to the same matrix. They then argue that, when analyzing θ_o , no adjustment need be made for the fact that λ_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 θ_o and studied predicting the median of y on the *original* data scale by backtransforming $x^t\hat{\theta}$. This idea of looking at the response surface avoids the problems of definition inherent with θ_o being defined in an unknown or data dependent scale. They found that when predicting the median of y, the effect of not knowing λ_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 θ_o.

The above discussion establishes the extent ot the controversy surrounding the Box and Cox model applied (1.5). We believe (1.4) entirely avoids this controversy. First, the parameter θ_o has physical meaning even if λ_o is unknown, since $f(x_i,\theta_o)$ 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 λ_o is estimated than when λ_o is known.

3: LIKELIHOOD ANALYSIS

The likelihood analysis proceeds as follows: define

$$z_i = dh(f_i(\theta_o),\lambda_o)/d\theta_o$$

$$f_i(\theta) = f(x_i,\theta),\ f_i = f_i(\theta_o),$$

$$h_y(y) = h_y(y,\lambda) = dh(y,\lambda)/dy, \text{ and } h(y) = h(y,\lambda).$$

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

as (all summations are from 1 to N)

(3.1)
$$N^{-1} I = \begin{pmatrix} S/\sigma_o^2 & 0 & C_1/\sigma_o^2 \\ \cdot & 1/(2\sigma_o^4) & C_2/\sigma_o^4 \\ \cdot & \cdot & C_3/\sigma_o^2 \end{pmatrix}$$

where

(3.2)
$$\begin{aligned} S &= N^{-1}\sum z_i z_i^t \\ C_1 &= -N^{-1}E\sum z_i[h_\lambda(y_i) - h_\lambda(f_i)]^t \\ C_2 &= -N^{-1}E\sum \epsilon_i[h_\lambda(y_i) - h_\lambda(f_i)]^t \\ C_3 &= N^{-1}E\sum\{[h_\lambda(y_i) - h_\lambda(f_i)][h_\lambda(y_i) - h_\lambda(f_i)^t \\ &\quad + \epsilon_i[h_{\lambda\lambda}(y_i) - h_{\lambda\lambda}(f_i)] + (\partial/\partial\lambda)(\partial/\partial\lambda)^t log[hy(y_i)]\}. \end{aligned}$$

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 $(\hat{\theta}, \hat{\sigma}^2, \hat{\lambda})^t$ is approximately $N((\theta_o \; \sigma_o^2 \; \lambda_o)^t, I^{-1})$ 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 λ_o is estimated differs from when λ_o 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 $f_i(\theta_o)$ 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 estimating λ_o, we follow Bickel and Doksum and others and let $\sigma_o \to 0$. While Bickel and Doksum let $N \to \infty$ and $\sigma_o \to 0$ simultaneously, we let $N \to \infty$ and then $\sigma_o \to 0$. 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 σ_o. In fact, the need for transformation is greater when σ_o is large. The small σ_o 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

$$C_1 = O(\sigma_o^2),\ C_2 = O(\sigma_o^2),\ \text{and } C_3 = O(\sigma_o^2) \text{ as } \sigma_o \to 0. \tag{3.3}$$

Standard calculations show that when λ_o is known,

$$N^{\frac{1}{2}} \text{ Covariance } [(\hat{\theta} - \theta_o)/\sigma_o,\ (\hat{\sigma}^2 \quad \sigma_o^2)/\sigma_o^2 | \lambda_o \text{ known}] \tag{3.4}$$
$$\to A^{-1} = \begin{pmatrix} S^{-1} & 0 \\ 0 & 2 \end{pmatrix}.$$

Let $D = Diag(\sigma_o,\ \sigma_o^2,\ 1)$. Then, to find this limiting covariance matrix when λ_o is unknown, we must find the upper left $(k+1) \times (k+1)$ corner of

$$DID = \begin{pmatrix} S & 0 & C_1/\sigma_o \\ \cdot & \frac{1}{2} & C_2/\sigma_o^2 \\ \cdot & \cdot & C_3/\sigma_o^2 \end{pmatrix}$$

which by standard results on inverting partitioned matrices is

$$A^{-1} + FE^{-1}F^t$$

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

$$E = C_3/\sigma_o^2 - B^t A\ B,$$
$$F = A^{-1}B,$$

and

$$B = \begin{pmatrix} C_1/\sigma_o \\ C_2/\sigma_o^2 \end{pmatrix} .$$

Clearly,

$$F = \begin{pmatrix} S^{-1}C_1/\sigma_o \\ 2C_2/\sigma_o \end{pmatrix}$$

and

$$E = C_3/\sigma_o^2 - C_1^t S^{-1} C_1/\sigma_o^2 - 2C_2^t C_2/\sigma_o^4 .$$

In order to obtain simple asymptotics, we will assume that for σ_o fixed, C_1/σ_o^2, C_2/σ_o^2, and C_3/σ_o^2 converge as $N \to \infty$, and that these, in turn, have limits D_1, D_2, *and* D_3 respectively as $\sigma_o \to 0$. 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

$$\lim_{\sigma_o \to 0} \lim_{N \to \infty} F\, E^{-1} F^t = \begin{pmatrix} 0 & 0 \\ 0 & W \end{pmatrix}$$

where $W = 4\, D_2^t\, [D_3 - D_2^t D_2]^{-1}\, D_2$. Therefore

$$\lim_{\sigma_o \to 0} \lim_{N \to \infty} (A^{-1} + F\, E^{-1} F^t) = \begin{pmatrix} S^{-1} & 0 \\ 0 & 2 + W \end{pmatrix}$$

<u>THEOREM 1</u>. Assume that the limits D_1 , D_2 , D_3 , S_∞ mentioned above exist and that $D_3 - 2D_2' D_2$ is nonsingular. As $N \to \infty$ and then $\sigma_o \to 0$, the limit distribution of $\hat{\theta}$ is the same whether λ_o is known or unknown. The limit distribution of $\hat{\sigma}$ depends on whether λ_o 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

$$h(y_i, \lambda) = x_i^t\, \theta_o + \epsilon$$

where $x_1, \dots, x_n$ are known $k \times 1$ vectors. Also, suppose that $\lambda_\circ = 0$, i.e., the log transformation is needed. Then $h_y(y) = y^{\lambda-1}$, $h_\lambda(y) = (log\ y)^2/2$, and $h_{\lambda\lambda}(y) = (log\ y)^3/3$. We find that

$$A = N^{-1}\sum x_i x_i^t/(x_i^t\theta_\circ)^2$$

$$C_1 = -(2N)^{-1}E\sum[x_i/(x_i^t\theta_\circ)]\{[log(x_i^t\theta_\circ) + \epsilon_i]^2-[log(x_i^t\theta_\circ)]^2\}$$
$$= -\sigma_\circ^2(2N)^{-1}\sum x_i/(x_i^t\theta_\circ),$$
$$C_2 = -(2N)^{-1}E\sum\epsilon_i\{[log(x_i^t\theta_\circ) + \epsilon_i]^2 - [log(x_i^t\theta_\circ)]^2\}$$
$$= -N^{-1}\sum log(x_i^t\theta_\circ)\ \sigma_\circ^2\ ,$$

and

$$C_3 = (4N)^{-1}E\sum\{[log(x_i^t\theta_\circ) + \epsilon_i]^2 - [log\ x_i^t\theta_\circ)]^2\}^2$$
$$+ (3N)^{-1}E\sum \epsilon_i\{[log(x_i^t\theta_\circ) + \epsilon_i]^3 - [log(x_i^t\theta_\circ)]^3\}$$
$$= 7/4\ \sigma_\circ^4 + 2\sigma_\circ^2/N \sum[log(x_i^t\theta_\circ)]^2\ .$$

Therefore,

$$D_1 = \lim_{N\to\infty}(2N)^{-1}\sum x_i/(x_i^t\theta_\circ)\ ,$$

$$D_2 = \lim_{N\to\infty} N^{-1}\sum log(x_i^t\theta_\circ),$$

and

$$D_3 = 2\lim_{N\to\infty} N^{-1}\sum[log(x_i^t\theta_\circ)^2,$$

provided the above limits exists. Thus, the 1×1 matrix $D_3 - 2D_2^tD_2$ is twice the limit of the variance of $log(x_1^t\theta_\circ), \dots, log(x_N^t\theta_\circ)$, 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 $\hat{\theta}(\hat{\lambda}_o)$ denote the ML estimator with λ_o estimated and known respectively. Let ARE$(\hat{\theta}_1,\hat{\theta}_2)$ be the asymptotic relative efficiency of $\hat{\theta}_1$ to $\hat{\theta}_2$. For fixed σ_o, it is difficult to find ARE$(\hat{\theta}(\lambda_o),\hat{\theta}(\hat{\lambda}))$ and, in fact, this may depend on θ_o, λ_o, the $\{x_i\}$ and the coordinate of θ 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 $\hat{\theta}(W)$and show that ARE$(\hat{\theta}(\lambda_o), \hat{\theta}(W)) \leq \pi/2$. Under reasonable regularity conditions, this means that 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 $w_1, \ldots , w_N$ be positive numbers and let $\theta(L)$ be any point which minimizes the expression

$$\sum w_i \; |y_i - f_i(\hat{\theta}(L))| \quad .$$

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

(4.1) $$0 \doteq \sum w_i \; sign \; (y_i - f_i(\hat{\theta}(L))s_i,$$
$$s_i = df_i(\theta_o)/d\theta.$$

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),

(4.2) $$0 \doteq \sum w_i \; \{sign(y_i - f_i(\hat{\theta}(L))) - sign(r_i)\}s_i$$
$$-(E\sum w_i\{sign(y_i - f_i(\theta)) - sign(r_i)\}s_i)| \; \theta=\hat{\theta}(L)).$$

Now, as $\epsilon \to 0$, we obtain that

(4.3) $$E(sign(r_i + \epsilon) - sign(r_i)) - 2\epsilon m_i(0) \to 0.$$

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

(4.4) $$(\hat{\theta}(L) - \theta_o) \doteq \tfrac{1}{2}(\sum w_i m_i(0) s_i s_i^{\,t})^{-1} \sum w_i s_i sign(r_i).$$

Now, since for model (1.4)

$$\epsilon_i = h(f_i(\theta_o) + r_i, \; \lambda_o) - h(f_i(\theta_o), \; \lambda_o),$$

we then have

(4.5) $$m_i(0) = (2\pi\sigma_o^2)^{-\frac{1}{2}} \; h_y(f_i(\theta_o), \; \lambda_o).$$

Thus, if we chose

$$(4.6) \qquad w_i = h_y(f_i(\theta_o), \lambda_o),$$

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

$$N^{\frac{1}{2}}(\hat{\theta}(L) - \theta_o)/\ \sigma_o \overset{L}{\to} N(0, (\pi/2)\ S^{-1}).$$

Now $\hat{\theta}(L)$ is not a bona fide estimator since w_i in (4.6) requires λ_o, θ_o to be known. However, if in (4.6) one plugs in any $N^{\frac{1}{2}}$ consistent estimators of θ_o and λ_o and calls the L_1 estimate based on these new weights $\hat{\theta}(W)$, then using Theorem 7.1 of Carroll and Ruppert (1982), one can also show that

$$N^{\frac{1}{2}}(\hat{\theta}(W) - \theta_o)/\sigma_o \overset{L}{\to} N(0, (\pi/2)S^{-1}).$$

Now, because

$$N^{\frac{1}{2}}(\hat{\theta}(\lambda_o) - \theta_o)/\sigma_o \overset{L}{\to} N(0,\ S^{-1}).$$

it then follows that

$$\text{ARE}(\hat{\theta}(\lambda_o),\ \hat{\theta}(W)) = \pi/2,$$

$$(4.8) \qquad \text{ARE}(\hat{\theta}(\lambda_o),\ \hat{\theta}(\hat{\lambda})) \le \pi/2.$$

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 σ_o.

The weighted L_1 estimator may well be useful for example if in (1.4) one suspected that the errors $\{\epsilon_i\}$ are not normal. It is a consistent estimator of θ_o provided that 0 is the unique median of ϵ_i. Symmetry of ϵ_i 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

$$(5.1) \qquad h(y_{ij},\lambda_\circ) = \mu_j + \epsilon_{ij} \quad j = 1, \ldots, k;\ i = 1,\ldots,N_i.$$

The equivalent formulation from our viewpoint is

$$(5.2) \qquad \begin{aligned} h(y_{ij},\lambda_\circ) &= h(\xi_j,\lambda_\circ) + \epsilon_{ij}, \\ \mu_j &= h(\xi_j,\lambda_\circ). \end{aligned}$$

Here ξ_j is the median of y_{ij} on the original scale and μ_j is the expected value of y_{ij} in the $\lambda_\circ$ scale. The results of Carroll and Ruppert imply that for estimating the ξ's, there is little cost in not knowing $\lambda_\circ$, while for estimating the μ's, Bickel and Doksum show that the cost of not knowing $\lambda_\circ$ can be enormous. Since

$$H_\circ:\mu_1 = \ldots = \mu_k \quad \text{iff} \quad H_\circ:\xi_1 = \ldots = \xi_k,$$

there should be little cost in testing for equality of means when $\lambda_\circ$ 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 $\sigma_\circ$ is not necessarily small, we undertook a small simulation of the model

$$(6.1) \qquad h(y_i,\lambda_\circ) = h(\theta_1 + \theta_2\, x_i,\lambda_\circ) + \sigma_\circ\epsilon_i,$$

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, λ_o known (KNOWN)
2) ML estimator, λ_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 θ_2 (for which it is consistent). We chose three values of σ_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 (λ_o, ϵ_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 (θ_1, θ_2) for each λ was done by the IMSL routine ZXSSQ while ZXGSN was used to estimate λ_o.

The results for the ML estimator with λ_o unknown (denoted MLE) are very encouraging. The mean square errors for MLE are quite close to those for KNOWN, the ML estimator with λ_o known, especially for the slope θ_2. These results agree with our small σ theory and indicate the minimal cost for not knowing λ_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(λ_o estimated)/KNOWN(λ_o 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 θ_2, the ML estimator MLE with λ_o unknown tends to dominate the LSE, especially for larger values of σ_o. In other words, for our model (1.4), there is real value to transformation when it is appropriate.

REFERENCES

BICKEL, PETER J. (1975). One step Huber estimates in the linear model. *J. Amer. Statist. Assoc. 70,* 428-433.

BICKEL, P.J., and DOKSUM, K.A. (1981). An analysis of transformations revisited. *J. Am. Statist. Assoc. 76,* 296-311.

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

BOX, G.E.P., and TIAO, G.C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Mass. Addison-Wesley.

CARROLL, R.J. (1980). A robust method for testing transformations to achieve approximate normality. *J. Roy. Statist. Soc. Series B 42,* 71-78.

CARROLL, R.J. (1982). Tests for regression parameters in power transformation models. Tentatively accepted by the *Scand. J. Statist.*

CARROLL, R.J., and RUPPERT, D. (1981). Prediction and the power transformation family. *Biometrika 68,* 609-617.

CARROLL, R.J. and RUPPERT, DAVID (1982). Robust estimation in heteroscedastic linear models. To appear in *Ann. Statist.*

DOKSUM, K.A., and WONG, C.W. (1981). *Statistical tests after transformations.* Manuscript.

HINKLEY, D.V., and RUNGER, G. (1981). Analysis of transformed data. To appear in *J. Am. Statist. Assoc.*

HOADLEY, B.A. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Ann. Math. Statist. 42,* 1977-1991.

RUPPERT, D. and CARROLL, R.J. (1980). Trimmed least squares estimation in the linear model. *J. Am. Statist. Assoc. 75,* 828-838.

SPITZER, J.J. (1978). A Monte-Carlo investigation of the Box-Cox transformation in small samples. *J. Am. Statist. Assoc. 73,* 488-495.

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 λ known.
MLE = ML estimate with λ unknown.
LSE = ordinary least squares estimate.

λ		0.00			0.25	
σ	0.05	0.10	0.50	0.05	0.10	0.50
BIAS OF KNOWN	0.03	0.06	0.56	0.01	0.03	0.23
MSE OF KNOWN	2.41	9.67	24.87	0.90	3.59	9.04
BIAS OF MLE	0.02	0.04	0.60	0.01	0.02	0.19
$\frac{\text{MSE OF MLE}}{\text{MSE OF KNOWN}}$	1.02	1.05	1.14	1.01	1.03	1.12
MSE OF MLE - MSE OF KNOWN	0.05	0.47	3.44	0.01	0.09	1.09
S.E. OF ABOVE DIFF.	0.02	0.15	0.77	0.01	0.04	0.25
BIAS OF LSE	0.11	0.40	9.48	0.04	0.13	2.60
$\frac{\text{MSE OF MLE}}{\text{MSE OF LSE}}$	0.97	0.90	0.22	1.00	0.98	0.63
MSE OF MLE - MSE OF LSE	-0.06	-1.15	-96.62	0.00	-0.06	-6.07
S.E. OF ABOVE DIFF.	0.04	0.33	4.71	0.01	0.06	0.78

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 λ   known.
MLE   = ML estimate with λ unknown.
LSE   = ordinary least squares estimate.

λ	0.00			0.25		
σ	0.05	0.10	0.50	0.05	0.10	0.50
BIAS OF KNOWN	0.01	0.01	0.03	0.00	0.01	0.02
MSE OF KNOWN	7.08	28.36	72.23	2.71	10.83	27.24
BIAS OF MLE	-0.01	-0.04	-0.15	0.00	-0.02	-0.16
$\frac{\text{MSE OF MLE}}{\text{MSE OF KNOWN}}$	1.06	1.06	1.01	1.06	1.06	1.03
MSE OF MLE - MSE OF KNOWN	0.41	1.57	0.95	0.15	0.60	0.72
S.E. OF DIFF.	0.10	0.40	0.67	0.04	0.77	0.27
BIAS OF LSE	0.05	0.15	2.97	0.02	0.04	0.50
$\frac{\text{MSE OF MLE}}{\text{MSE OF LSE}}$	0.98	0.96	0.59	1.01	1.01	0.91
MSE OF MLE - MSE OF LSE	-0.16	-1.29	-50.54	0.05	0.13	-2.81
S.E. OF DIFF.	0.18	0.80	5.10	0.06	0.23	0.74

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