The Retransformed Mean After a Fitted Power Transformation

JEREMY M. G. TAYLOR*

An approximate method is given to estimate the mean of a dependent variable after a linear model is fitted to the Box–Cox power transformation of this variable. The estimate is accurate except when the transformation power parameter is near zero. The properties of the estimate in the one-sample and regression cases are considered, by both asymptotic calculations and Monte Carlo simulations, and comparisons are made with the smearing estimate (Duan 1983). It is shown that there can be some cost due to estimating the power transformation, as opposed to assuming it is known; however, this cost is not severe.

KEY WORDS: Box–Cox transformation; Mean prediction; Smearing estimate; Small-θ approximation.

1. INTRODUCTION

Box and Cox (1964) discussed the power transformation family of models. In particular, they gave methods of estimating the parameters in the nonlinear model

$$h(Y, \lambda) = \mathbf{A}\,\boldsymbol{\beta} + \sigma e, \tag{1}$$

where e has density f, which is assumed to be standard normal, $\mathbf{A}$ is a known design matrix, $\boldsymbol{\beta}$ is a vector of parameters, and

$$\begin{aligned} h(Y, \lambda) &= (Y^{\lambda} - 1)/\lambda, \qquad && \lambda \neq 0 \\ &= \log(Y), && \lambda = 0. \end{aligned}$$

Note that the transformation is valid only if $Y > 0$.

Equation (1) is optimistically assuming that a single transformation can achieve a linear structure, constant variance, and normal errors. In practice this will rarely be true; however, the exactness of (1) may not be important.

There are many methods of estimating the parameters; the one considered here is by maximum likelihood (Box and Cox 1964). The parameter estimates are consistent if f is normal; otherwise they are generally inconsistent, but the bias in the estimate of λ is small (Taylor 1985a).

Recently (Bickel and Doksum 1981; Box and Cox 1982; Hinkley and Runger 1984) there has been some discussion about the means and appropriateness of inferences concerning $\boldsymbol{\beta}$. In particular, Bickel and Doksum showed that the variance of $\hat{\boldsymbol{\beta}}$ is greatly inflated when λ is estimated from the data compared to the situation in which λ is assumed known. Carroll and Ruppert (1981) argued for parameters defined independently of the scale. They suggested using the inverse transformation and making inference statements on the original scale of the variables. They studied the properties of the conditional median of the distribution of Y given $\mathbf{A}$—that is, the retransformed median after a liner model is fitted to the transformed observations. They showed that there is some cost due to estimating λ but that it is generally small. Furthermore, it is much smaller than the cost obtained by Bickel and Doksum for the estimates of $\boldsymbol{\beta}$.

If one wishes to predict a future value of Y given $\mathbf{A}$, then on the original scale of the variables the conditional mean will frequently be the quantity of interest (see Morris 1984). For example, if the dependent variable is a monetary variable, then the mean value rather than the median value will probably be more relevant. Here an approximate method of estimating the conditional mean is given, its properties are studied, and it is compared to the smearing estimate (see Duan 1983 and Carroll and Ruppert 1984). The approximation uses the "small-θ" method (Draper and Cox 1969; Taylor 1985a). The two estimators are studied in the one-sample and regression cases, using both asymptotic calculations and Monte Carlo simulations. Except in the region of the parameter space where λ is near zero and σ is large, it is shown that the approximate method is very accurate and that the smearing estimate and the approximation are essentially equivalent. Further, I show that there is a cost due to estimating λ, compared to the λ-known situation, but this cost is generally small. For the examples considered in this article, the cost is of the order of roughly 30% or less. This was the conclusion reached by Carroll and Ruppert (1981) for the conditional median and suggested for the conditional mean.

2. APPROXIMATE ESTIMATOR

Assume that $Y_1, \ldots, Y_n$ are generated according to (1), with f a general density satisfying $E_f(e) = 0$ and $E_f(e^2) = 1$. Then the conditional mean of a future value of Y when $\mathbf{A} = \mathbf{A}_0$ is given by

$$E(Y|\mathbf{A}_0) = \int (1 + \lambda \mathbf{A}_0 \mathbf{B} + \lambda \sigma e)^{1/\lambda}\, dF(e).$$

An obvious estimate of this quantity is

$$\int (1 + \hat{\lambda}\, \mathbf{A}_0 \hat{\boldsymbol{\beta}} + \hat{\lambda} \hat{\sigma} e)^{1/\hat{\lambda}}\, dF(e). \tag{2}$$

Evaluating this would require estimating the parameters $\xi = (\lambda, \boldsymbol{\beta}, \sigma)$, estimating the density f, and evaluating the integral numerically—clearly not a very satisfactory method.

To obtain an approximation to this quantity, I use the small-θ method given in Draper and Cox (1969). Let $\theta(\mathbf{A}_0) = \lambda\sigma/(1 + \lambda \mathbf{A}_0 \boldsymbol{\beta})$, $\lambda \neq 0$. Then the condition

$$|\theta(\mathbf{A}_0)| \ll 1 \text{ for all } \mathbf{A}_0 \text{ of interest}$$

is necessary to ensure that all of the Y observations are positive

* Jeremy M. G. Taylor is Assistant Professor in Residence, Division of Biostatistics, School of Public Health, and with the Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90024. This work was supported by National Cancer Institute Grant CA 16042. The author thanks the associate editor for helpful suggestions.

© 1986 American Statistical Association
Journal of the American Statistical Association
March 1986, Vol. 81, No. 393, Theory and Methods

with high probability. So

$$E(Y|\mathbf{A}_0) = (1 + \lambda\mathbf{A}_0\boldsymbol{\beta})^{1/\lambda} \int (1 + \theta(\mathbf{A}_0)e)^{1/\lambda}\, dF(e)$$
$$= (1 + \lambda\mathbf{A}_0\boldsymbol{\beta})^{1/\lambda}\, I.$$

Expanding the integrand in powers of $\lambda(\mathbf{A}_0)$ gives

$$I = \int \left\{1 + \sum_{i=1}^{\infty} (\theta(\mathbf{A}_0)e)^i \prod_{j=1}^{i-1} \left(\frac{1}{\lambda} - j\right)\right\} dF(e).$$

If the density f is symmetric and the order of integration and summation can be interchanged (conditions that are similar to those given in Taylor 1985b), then deleting terms of a higher order than $(\theta(\mathbf{A}_0))^2$ gives

$$E(Y|\mathbf{A}_0) \simeq Y_a = (1 + \lambda\mathbf{A}_0\boldsymbol{\beta})^{1/\lambda}\left\{1 + \frac{\sigma^2(1 - \lambda)}{2(1 + \lambda\mathbf{A}_0\boldsymbol{\beta})^2}\right\}.$$

So an obvious estimator to use is

$$\hat{Y}_a = (1 + \hat{\lambda}\mathbf{A}_0\hat{\boldsymbol{\beta}})^{1/\hat{\lambda}}\{1 + \hat{\sigma}^2(1 - \hat{\lambda})/2(1 + \hat{\lambda}\mathbf{A}_0\hat{\boldsymbol{\beta}})^2\}.$$

Note the following: (a) $\hat{Y}_a$ is equivariant to changes in the Y scale and (b) Y_a can be viewed as the median multiplied by a correction factor, where the correction factor is larger than 1 if $\lambda < 1$ and less than 1 if $\lambda > 1$. See Morris (1984) for a similar bias correction when $\lambda = 1/N$.

Two questions need to be addressed. First, how close is Y_a to $E(Y|\mathbf{A}_0)$? Clearly for λ near zero, $Y_a \simeq \exp(\mathbf{A}_0\boldsymbol{\beta})(1 + \sigma^2/2)$; however, in this case if f is normal, then $E(Y|\mathbf{A}_0) \simeq \exp(\mathbf{A}_0\boldsymbol{\beta} + \sigma^2/2)$. Therefore the approximation is not accurate in the situation where λ is near zero and σ is large. I will vaguely define this region of the parameter space to be S' and its complement to be S; that is, $S' = \{\xi : \lambda \text{ near zero and } \sigma \text{ large}\}$, where $\xi = (\lambda, \boldsymbol{\beta}, \sigma)$. Unfortunately S' is an important region, but as we shall see it is also a region where it is very hard to get a reasonable estimator. In Section 4, I show that S' appears to be the only region where the approximation is not valid.

Second, what are the properties of $\hat{Y}_a$ as an estimator of Y_a? A simple Taylor expansion gives

$$\hat{Y}_a = Y_a + P(\hat{\lambda} - \lambda) + \mathbf{Q}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$
$$+ R(\hat{\sigma} - \sigma) + o(\hat{\lambda} - \lambda, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \hat{\sigma} - \sigma). \quad (3)$$

So if the parameters ξ are consistently estimated, then $\hat{Y}_a \xrightarrow{p} Y_a$ as $n \rightarrow \infty$.

Similarly, the limiting variance of $n^{1/2}(\hat{Y}_a - Y_a)$ can be obtained from the asymptotic covariance of ξ. I do not give its form here because the terms in the matrix involve complicated integrals (although approximate answers can be obtained in simple cases) and the terms P, $\mathbf{Q}$, and R are lengthy expressions.

3. SMEARING ESTIMATE

Duan (1983) described a nonparametric method of estimating the conditional mean when the data follow a linear model after a known transformation. He calls the method the smearing estimate.

Duan showed that the estimator is consistent and, in the case of the lognormal distribution, that the loss in efficiency relative to the normal theory estimate is small unless σ is large. The bias of the normal theory estimate, however, is sensitive to departures from normality. In the one-sample case the smearing estimate reduces to the sample mean.

Carroll and Ruppert (1984) suggested using the smearing estimate when the transformation is estimated from the data. The method is to replace F in (2) by the empirical distribution function of the residuals $\hat{e}_i = (h(Y_i, \hat{\lambda}) - (\mathbf{A}\hat{\boldsymbol{\beta}})_i)/\hat{\sigma}$. This leads to the estimator

$$\hat{Y}_S = \frac{1}{n} \sum (Y_i^{\hat{\lambda}} + \hat{\lambda}((\mathbf{A}_0 - \mathbf{A})\hat{\boldsymbol{\beta}})_i)^{1/\hat{\lambda}}, \qquad \hat{\lambda} \neq 0$$
$$= \frac{1}{n} \sum Y_i \exp\{((\mathbf{A}_0 - \mathbf{A})\hat{\boldsymbol{\beta}})_i\}, \qquad \hat{\lambda} = 0.$$

The theoretical properties of this estimator are being discussed elsewhere (Duan, personal communication, 1985).

Proposition 1. The approximate estimator and the smearing estimator are approximately equal except when λ is near zero.

Proof.

$$\hat{Y}_s = \frac{1}{n} \sum (1 + \hat{\lambda}(\mathbf{A}_0\hat{\boldsymbol{\beta}} + \hat{\sigma}\hat{e}_i))^{1/\hat{\lambda}}$$
$$= (1 + \hat{\lambda}\mathbf{A}_0\boldsymbol{\beta})^{1/\hat{\lambda}} \frac{1}{n} \sum (1 + \hat{\theta}\hat{e}_i)^{1/\hat{\lambda}} \text{ for } \hat{\lambda} \neq 0,$$

where $\hat{\theta} = \hat{\lambda}\hat{\sigma}/(1 + \hat{\lambda}\mathbf{A}_0\hat{\boldsymbol{\beta}})$, which is assumed small. Hence

$$\hat{Y}_S = (1 + \hat{\lambda}\mathbf{A}_0\hat{\boldsymbol{\beta}})^{1/\hat{\lambda}} \frac{1}{n} \sum \left\{1 + \frac{\hat{\theta}\hat{e}_i}{\hat{\lambda}}\right.$$
$$\left. + \frac{\hat{\theta}^2\hat{e}_i^2}{2} \frac{1}{\hat{\lambda}} \left(\frac{1}{\hat{\lambda}} - 1\right) + 0(\hat{\theta}^3)\right\}.$$

The normal equations give $\sum \hat{e}_i = 0$, provided the linear model contains an intercept term, and $\sum \hat{e}_i^2 = n$. Hence

$$\hat{Y}_S \simeq (1 + \hat{\lambda}\mathbf{A}_0\hat{\boldsymbol{\beta}})^{1/\hat{\lambda}}(1 + (\hat{\theta}^2/2)(1/\hat{\lambda})(1/\hat{\lambda} - 1)) = \hat{Y}_a.$$

This result is borne out in the Monte Carlo simulations.

4. ONE-SAMPLE CASE

4.1 Theoretical Results

Assume the model

$$h(Y, \lambda) = \mu + \sigma e, \qquad (4)$$

where e has density f, which is normal. The mean of Y is given by

$$E(Y) = \int (1 + \lambda\mu + \lambda\sigma e)^{1/\lambda} f(e)\, de.$$

Estimating this quantity may be a hard task. For example, if $\lambda = 0$, then a small departure from the lognormal distribution in the right-hand tail can greatly affect $E(Y)$. In the sampling situation, whether or not one of the extreme observations is sampled will play an enormous role in determining the estimate of $E(Y)$.

The smearing estimator $\hat{Y}_S$ reduces to the sample mean, $\overline{Y}$, the properties of which are easy to assess.

The small-θ estimator is

$$\hat{Y}_a = (1 + \hat{\lambda}\hat{\mu})^{1/\hat{\lambda}}(1 + \hat{\sigma}^2(1 - \hat{\lambda})/2(1 + \hat{\lambda}\hat{\mu})^2),$$

where the parameter estimates are obtained using the maximum

Table 1. Comparison of the Limiting Value of the Approximate Estimator (Y^*) and E(Y) When the Distribution of the Data Is Gamma (α, 1)

Shape Parameter α	Y^*	$E(Y)$	λ^*	σ^*	θ^*
.5	.475	.5	.208	1.27	.36
1.0	.994	1.0	.265	1.01	.30
1.5	1.498	1.5	.289	.92	.25
2.0	2.000	2.0	.301	.87	.22
3.0	2.999	3.0	.312	.81	.19

likelihood method. For general f, $\hat{Y}_a$ is estimating the quantity

$$Y^* = (1 + \lambda^*\mu^*)^{1/\lambda^*}(1 + \sigma^{*2})/2(1 + \lambda^*\mu^*)^2),$$

where $\xi^* = (\lambda^*, \mu^*, \sigma^*)$ are the limiting values of $\hat{\xi}$. If model (4) is the true model, then $\xi^* = \xi$, and in general ξ^* is evaluated using the method of Hernandez and Johnson (1980).

Table 1 gives a comparison of Y^* and $E(Y)$ when the actual distribution of the data is gamma (α, 1). As expected, we see that Y^* is close to $E(Y)$ except when λ is near zero and σ is large. The values of $\theta^* = \lambda^*\sigma^*/(1 + \lambda^*\mu^*)$ are also given.

I further compared the approximation Y_a, with $E(Y)$ calculated numerically, assuming that model (4) is true, for many configurations of the parameters $0 \le \lambda \le 2$. [For $\lambda < 0$, $E(Y)$ can be infinite.] The two values were found to be very close to each other for all ξ, except for both λ near zero ($\lambda < .2$) and σ large ($\sigma > 1$).

Table 2 gives the asymptotic relative bias of Y_a when f is assumed to be normal and contaminated normal [$N(0, 1)$ with probability .9 and $N(0, 9)$ with probability .1, standardized to have unit variance]. The table is given in terms of λ and θ, the more natural parameters in the one-sample case. Note that θ ($= \lambda\sigma/(1 + \lambda\mu)$) is the coefficient of variation of Y and that θ large is approximately equivalent to σ large. Again we see that the bias is negligible unless $\xi \in S'$.

It is easy to show that, under model (4), the limiting variance (Avar) of $n^{1/2}\bar{Y}$ equals $\sigma^2(1 + \lambda\mu)^{(2/\lambda)-2}$. Similarly, after a lengthy likelihood calculation involving a small-θ approximation to the covariance matrix of $\hat{\xi}$, it can be shown that for $\xi \in S$ and f normal,

$$\text{Avar}(n^{1/2}\hat{Y}_a) = \sigma^2(1 + \lambda\mu)^{(2/\lambda)-2}(1 + 0(\theta^2)). \qquad (5)$$

Table 2. Asymptotic Relative Bias of $\hat{Y}_{aj}$ (100 × (Y_a^* − E(Y))/E(Y))

			Density	
λ	θ	σ	Normal	Contaminated Normal
1.0	all θ		.0	.0
.5	all θ		.0	.0
.25	.1		.0	−.1
	0.2		−.4	−1.1
	0.3		−1.6	*
.1	0.1		−4.4	−13.1
	0.2		−30.4	−68.7
	0.3		−60.4	*
.0		.5	−.1	−2.3
		1.0	−9.0	−37.7
		1.5	−31.0	−92.8

* Not a permissible value of θ for this density.

This is hardly a surprising result, as we knew that $\hat{Y}_S$ and $\hat{Y}_a$ were approximately equal.

Proposition 2. When estimating the mean on the original scale of the observations, there is no additional cost incurred [to $0(\theta^2)$] due to estimating the power transformation, assuming that model (4) is true with f normal and $\xi \in S$.

Proof. If λ is assumed known, then a simple likelihood calculation shows that

$$\text{Avar}(n^{1/2}\hat{Y}_a|\lambda \text{ known}) = \sigma^2(1 + \lambda\mu)^{(2/\lambda)-2}(1 + 0(\theta^2)).$$

Comparison with Equation (5) gives the result.

4.2 Monte Carlo Results

Five hundred samples of size 30 were generated according to model (4), for a variety of parameter configurations. Normal and contaminated-normal densities were considered. Strictly speaking, truncated densities were considered. All nonpositive values of Y generated in Equation (4) were discarded. The effect of this was thought to be negligible, as less than .1% of the generated Y values were discarded. Table 3 gives the results.

The table shows that for $\xi \in S$, $\hat{Y}_a$ and $\bar{Y}$ have essentially equal variance when f is normal, but $\hat{Y}_a$ is slightly more efficient when f is contaminated normal. As expected, the variance ratio, $\text{var}(\hat{Y}_a)/\text{var}(\hat{Y}_a|\lambda \text{ known})$, is close to one when f is normal and less than one if f is contaminated normal, showing that there is essentially no cost due to estimating λ. The results concerning the bias of $\hat{Y}_a$ are not shown; however, in all cases (except $\xi \in S'$) the difference between the Monte Carlo means of $\hat{Y}_a$ and $\bar{Y}$ was negligible, and the relative bias was very similar to the values given in Table 2.

When $\xi \in S'$, $\hat{Y}_a$ is more efficient than $\bar{Y}$, but this is also the region where the bias of $\hat{Y}_a$ becomes significant. So it seems that in S', $\bar{Y}$ retains zero bias at the expense of variance and $\hat{Y}_a$ retains low variance at the expense of bias. If the mean squared error criterion is used, then $\hat{Y}_a$ is better than $\bar{Y}$, at least for samples of size 30. Despite this smaller mean squared error, it would not be sensible to use $\hat{Y}_a$ if the bias is large; however, if one is prepared to tolerate a 5% bias (say), then $\hat{Y}_a$ can be up to twice as efficient as $\bar{Y}$.

Table 3. One-Sample Monte Carlo Results: Efficiency of $\hat{Y}_a$ to $\bar{Y}$ and Efficiency of $\hat{Y}_a$ to ($\hat{Y}_a|\lambda$ known)

			$\text{var}(\bar{Y})/\text{var}(\hat{Y}_a)$		$\text{var}(\hat{Y}_a)/\text{var}(\hat{Y}_a\|\lambda \text{ known})$	
λ	$\|\theta\|$	σ	Normal	Contaminated Normal	Normal	Contaminated Normal
1.0	.1		1.01	1.12	.99	.89
1.0	.2		1.00	1.13	1.00	.89
1.0	.3		1.00		1.00	
.5	.1		1.00	1.12	1.00	.90
.5	.2		1.00	1.16	1.00	.86
.5	.3		1.00		1.00	
.25	.1		1.01	1.27	1.00	.87
.25	.2		1.03	2.12	.99	.65
.25	.3		1.04		.99	
−.5	.1		1.09	3.81	.99	.78
−.5	.2		1.75		.97	
−1.0	.1		1.01	1.01	1.00	.77
−1.0	.2		1.13	.99		
.0		.5	1.14	3.12	.97	.76
.0		1.0	1.77	5 × 10^3	.99	.53
.0		1.5	6.95	3 × 10^4	.98	.47

5. LINEAR REGRESSION

5.1 Theoretical Results

Consider the model $h(Y_i, \lambda) = \beta_0 + \beta_1 C_1 + \sigma e_i$. Without loss of generality, restrict $\{C_i\}$ such that $\sum_{i=1}^n C_i = 0$ and $\sum_{i=1}^n C_i^2 = n$. At a value C_0 of the independent variable, the smearing estimate is given by

$$\hat{Y}_S(C_0) = \frac{1}{n}\sum_{i=1}^{n} (Y_i^{\hat{\lambda}} + \hat{\lambda}\hat{\beta}_1(C_0 - C_i))^{1/\hat{\lambda}}$$

and the small-θ approximation is given by

$$\hat{Y}_a(C_0) = R^{1/\hat{\lambda}}(1 + \tfrac{1}{2}\hat{\sigma}^2(1 - \hat{\lambda})/R^2),$$

where $R = 1 + \hat{\lambda}(\hat{\beta}_0 + \hat{\beta}_1 C_0)$.

For a normal error distribution we can in theory obtain the asymptotic covariance matrix of ξ, but in practice this would involve numerical integrals. A small-σ approximation to this matrix was given by Bickel and Doksum (1981); also see Carroll and Ruppert (1981). Then

$$\frac{n}{\sigma^2}\operatorname{cov}(\hat{\lambda}, \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}) = \begin{pmatrix} \Sigma & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \tag{6}$$

where

$$\Sigma = e^{-1}\begin{pmatrix} 1 & -E(T) & -E(CT) \\ -E(T) & E(T^2) - (E(CT))^2 & E(T)E(CT) \\ -E(CT) & E(T)E(CT) & E(T^2) - (E(T))^2 \end{pmatrix},$$

where $T = (1/\lambda^2)(S - 1 - S\log(S))$, $S = 1 + \lambda(\beta_0 + \beta_1 C)$, $e = E(T^2) - (E(T))^2 - (E(CT))^2$, and the expectations are with respect to the distribution of C.

It is worth noting that the variances given by this approximation were in good agreement with the Monte Carlo results when σ was small; however, there were substantial differences between the two when σ was not small. Carroll and Ruppert (1981) noted a similar discrepancy.

By using (3) and (6) we obtain the limiting variance of $\hat{Y}_a(C_0)$. In a similar manner we can obtain the limiting variance of the approximate estimator, assuming λ is known. These variances and their ratio are given in Table 4 for a particular case.

Table 4. Linear Regression Results

	Relative Bias $(100 \times (\hat{Y} - Y^*)/Y^*)$, Monte Carlo		Variance Ratio $(\text{var}(\hat{Y})/\text{var}(\hat{Y}\mid\lambda \text{ known}))$		
			Monte Carlo		Small-σ Analysis,
C_0	$\hat{Y}_S(C_0)$	$\hat{Y}_a(C_0)$	$\hat{Y}_S(C_0)$	$\hat{Y}_a(C_0)$	$\hat{Y}_a(C_0)$
−1.508	.0053	.0055	1.22	1.22	3.38
−1.206	.0066	.0068	1.04	1.04	2.09
−.905	.0061	.0061	1.02	1.02	1.33
−.603	.0046	.0046	1.14	1.14	1.02
−.302	.0029	.0029	1.32	1.32	1.02
.302	−.0001	−.0001	1.28	1.28	1.02
.603	−.0008	−.0008	1.10	1.10	1.03
.905	−.0006	−.0007	1.00	1.00	1.32
1.206	.00053	.00045	1.04	1.04	1.94
1.508	.0028	.0027	1.22	1.21	2.86

NOTE: $\lambda = .5, \beta_0 = 4, \beta_1 = 1, \sigma = .5; f = \text{normal}, n = 20$.

Table 5. Monte Carlo Regression Results

		Relative Difference	Efficiency	$(\text{var}(Y)/\text{var}(Y\mid\lambda$ known))	
Density	σ	$(100 \times (\hat{Y}_S - \hat{Y}_a)/\hat{Y}_S)$	$(\text{var}(\hat{Y}_a)/\text{var}(\hat{Y}_S))$	$\hat{Y}_a$	$\hat{Y}_S$
Normal	.5	1	.96	1.11	1.12
	1.0	10	.63	1.04	1.43
Contaminated	.5	4	.50	1.32	2.4
Normal	1.0	~50	.00	.80	~100

NOTE: $\lambda = .0, \beta_0 = -3, \beta_1 = 1$. The numbers in this table are averages over the 10 values of C_0.

5.2 Monte Carlo Results

Five hundred samples of size 20 were generated for a range of values of ξ ($\lambda = 1.0, .5, .25, .0, -1.0$ and $\sigma = .5, 1.0$). Normal and contaminated normal errors were considered. The values of $\{C_i\}$ were always those given in Table 4, with two observations at each value. The Monte Carlo mean and variance of the approximate estimator and the smearing estimator were calculated for λ estimated and assumed known. Table 4 gives the detailed results for one particular case, and Table 5 summarizes the results when $\lambda = .0$.

Table 4 shows that the two estimators are essentially equivalent and have small bias and that the cost due to estimating λ is fairly small. In addition, the small-σ asymptotics provide a reasonable approximation in terms of order of magnitude, but are not particularly accurate.

The general findings of the Monte Carlo study are as follows:

1. When f is normal and $\xi \in S$, $\hat{Y}_a(C_0)$ and $\hat{Y}_S(C_0)$ are essentially equivalent, the cost due to estimating λ is minimal (seldom greater than 30%), and the relative bias of $\hat{Y}_a(C_0)$ compared to $\hat{Y}_S(C_0)$ is negligible.
2. When f is contaminated normal and $\xi \in S$, $\hat{Y}_a(C_0)$ is slightly more efficient than $\hat{Y}_S(C_0)$ (0%–20%); there is no cost due to estimating λ (in fact in some cases there is something to be gained). The relative bias of $\hat{Y}_a(C_0)$ compared to $\hat{Y}_S(C_0)$ is small (<2%).
3. When λ is near zero and σ is large ($\sigma \geq 1$), then $\hat{Y}_a(C_0)$ is more efficient than $\hat{Y}_S(C_0)$ (extremely so the larger σ gets or if f is contaminated normal). The bias of $\hat{Y}_a(C_0)$ compared to $\hat{Y}_S(C_0)$ can be large. There is a small cost due to estimating λ when f is normal and a small gain when f is contaminated normal.
4. The cost due to estimating λ is not constant throughout the design space. This was also noted by Carroll and Ruppert (1984) for the conditional median.

The precise results for $\lambda = 0$ are given in Table 5. The numbers given are averaged over the 10 values of C_i. The table illustrates the extreme problem in estimating the mean if $\lambda = 0$ and σ is large. One estimator has a very large bias, whereas the other estimator has a very large variance.

6. CONCLUSIONS

Two methods are given of estimating the conditional mean after a transformation is fitted to the data; neither method requires numerical integration or density estimation. Except for both λ near zero and σ large, the two methods are essentially equivalent. They are estimating the correct quantity, and there is generally only a small to moderate cost incurred due to estimating λ. This cost may or may not be important depending on the context and which part of the design space is of interest. This small cost suggests that when estimating the conditional mean it may be satisfactory, as a first approximation, to use the estimated value of λ as if it was known beforehand. In a final analysis, however, more accurate confidence intervals for the conditional mean could be obtained by using a bootstrap technique.

When λ is near zero and σ is large, the problem of estimating the conditional mean is much harder. The two methods differ here; the small-θ approximation method retains small efficiency at the expense of bias, whereas the smearing estimate retains zero bias at the expense of poor efficiency.

In this problem area, the bias of the small-θ estimator could be reduced by retaining higher-order terms in the expansion of $E(Y|\mathbf{A}_0)$ in powers of $\theta(\mathbf{A}_0)$. Alternatively, making the smearing estimator robust would reduce its variance.

[*Received March 1985. Revised August 1985.*]

REFERENCES

Bickel, P. J., and Doksum, K. A. (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association,* 76, 296–311.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society,* Ser. B, 26, 211–252.

——— (1982), "An Analysis of Transformations Revisited, Rebutted," *Journal of the American Statistical Association,* 77, 209–210.

Carroll, R. J., and Ruppert, D. (1981), "Prediction and the Power Transformation Family," *Biometrika,* 68, 609–616.

——— (1984), Comment on "The Analysis of Transformed Data," by D. V. Hinkley and G. Runger, *Journal of the American Statistical Association,* 79, 312–313.

Draper, N. R., and Cox, D. R. (1969), "On Distributions and Their Transformations to Normality," *Journal of the Royal Statistical Society,* Ser. B, 31, 472–476.

Duan, N. (1983), "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association,* 78, 605–610.

Hernandez, F., and Johnson, R. (1980), "The Large-Sample Behavior of Transformations to Normality," *Journal of the American Statistical Association,* 75, 855–861.

Hinkley, D. V., and Runger, G. (1984), "The Analysis of Transformed Data," *Journal of the American Statistical Association,* 79, 302–309.

Morris, D. M. (1984), "Reducing Transformation Bias in Curve Fitting," *The American Statistician,* 38, 124–126.

Taylor, J. M. G. (1985a), "Power Transformations to Symmetry," *Biometrika,* 72, 145–152.

——— (1985b), "Measures of Location of Skew Distributions Obtained Through Box–Cox Transformations," *Journal of the American Statistical Association,* 80, 427–432.