ON PREDICTION AND THE POWER TRANSFORMATION FAMILY

by

R.J. Carroll* and David Ruppert**
University of North Carolina

Abstract

The power transformation family studied by Box and Cox (1964) for transforming to a normal linear model has recently been further studied by Bickel and Doksum (1978), who show that "the cost of not knowing λ and estimating it . . . is generally severe"; some of the variances for regression parameters can be extremely large. We consider prediction of future observations (untransformed) when the data can be transformed to a linear model and show that while there is a cost due to estimating λ it is generally not severe. Similar results emerge for the two sample problems. Monte-Carlo results lend credence to the asymptotic calculations.

Key Words and Phrases: Power transformations, Prediction, Robustness, Box-Cox family, Asymptotic theory.

AMS 1970 Subject Classification: Primary 62E20; Secondary 62G35.

---

*Research for this work was supported by the Air Force Office of Scientific Research under Grants AFOSR-75-2796 and AFOSR-80-0080.

**Research for this work was supported by the National Science Foundation under Grant MCS78-01240.

1. Introduction

The power transformation family studied by Box and Cox (1964) takes the following form: for some unknown λ,

$$(1.1) \qquad y_i^{(\lambda)} = \underline{x}_i \underline{\beta} + \varepsilon_i \qquad (i = 1,\ldots,N) ,$$

where $\{\varepsilon_i\}$ are independently and identically distributed with distribution F and where

$$y^{(\lambda)} = (y^{\lambda}-1)/\lambda \qquad \lambda \neq 0$$
$$= \log y \qquad \lambda = 0 .$$

Box and Cox propose maximum likelihood estimates for λ and $\underline{\beta}$ when F is the normal distribution. There are numerous alternate methods as well as proposals for testing hypotheses of the form H_0: $\lambda = \lambda_0$ (Hinkley (1975), Andrews (1971), Atkinson (1973), Carroll (1978)). Carroll studied the testing problem via Monte-Carlo; by allowing F to be non-normal he approximated a problem with outliers and found that the chance of mistakenly rejecting the null hypothesis can be very high indeed.

Bickel and Doksum (1978, hereafter denoted B-D) develop an asymptotic theory for estimation. They assume that $\underline{x}_1, \underline{x}_2, \ldots$ are independent and identically distributed according to G, a lead which we follow throughout. If the maximum likelihood estimate of the regression parameter is $\hat{\underline{\beta}}(\hat{\underline{\beta}}(\hat{\lambda}))$ when λ is known (unknown and estimated by $\hat{\lambda}$), they compute the asymptotic distributions of $n^{\frac{1}{2}}(\hat{\underline{\beta}} - \underline{\beta})/\sigma$ and $n^{\frac{1}{2}}(\hat{\underline{\beta}}(\hat{\lambda}) - \underline{\beta})/\sigma$ as $n\to\infty, \sigma\to 0$. These distributions are different and as regards variances:

> 'the cost of not knowing λ and estimating it . . . is generally severe. . . . The problem is that $(\hat{\underline{\beta}}(\hat{\lambda}))$ and $\hat{\lambda}$ are highly correlated. Thus, the assertion (of Box and Cox) that "having chosen a suitable λ, we should make the usual detailed examination and interpretation on this transformed scale" is incorrect.'

They thus show that $\hat{\lambda}$ and $\hat{\underline{\beta}}(\hat{\lambda})$ are unstable (i.e., highly variable) and highly correlated, a problem similar in nature to that of multicollinearity in regression.

These conclusions made by B-D seem to have caused some controversy. One point of discussion concerns the scale on which inference is to be made: i.e., should one make unconditional inference about the regression parameter in the correct but unknown scale (the B-D theory) or a conditional inference for an appropriately defined "regression parameter" in an estimated scale?

In order to eliminate such problems, we will study the cost of estimating λ when one wants to make inferences in the *original* scale of the observations. In the multicollinearity problem, reasonably good prediction is still possible if new vectors $\underline{x}$ arrive independently with the distribution G. Motivated by this fact (see the discussion in Section 2), we will focus our attention specifically on prediction, but we also discuss the 2-sample problem and a somewhat more general estimation theory. Using the B-D asymptotic theory and Monte-Carlo, we find that for prediction (and other problems in the original scale), there is a cost due to estimating λ, but it is generally *not* severe.

2. Predicting the conditional median in regression

Our model specifically includes an intercept, i.e., $\underline{x}_i = (1,\underline{c}_i)$; by suitable rescaling we assume the $\underline{c}_i$ have mean zero and identity covariance. From the sample we calculate $\hat{\lambda}$ and $\hat{\underline{\beta}}(\hat{\lambda})$, and we are given a new vector

$\underline{x}_0 = (1,\underline{c}_0)$ which is independent of the other $\underline{x}$'s but still has the same distribution G. Our predicted value in the transformed scale would be $\underline{x}_0\hat{\underline{\beta}}(\hat{\lambda})$, so a natural predictor is

$$f(\hat{\lambda},\underline{x}_0\hat{\underline{\beta}}(\hat{\lambda})) \quad , \tag{2.1}$$

where

$$\begin{aligned} f(\lambda,\theta) &= (1+\lambda\theta)^{1/\lambda} \qquad \lambda \neq 0 \\ &= \exp(\theta) \qquad\quad \lambda = 0 \ . \end{aligned}$$

Notice that if F is symmetric, or more generally has median equal to 0, then $f(\lambda,\underline{x}_0\beta)$ is the median of the conditional distribution of y given x_0, even though it is not necessarily the conditional expectation. Calculation of conditional expectations would require the use of numerical integration and that F be known or an estimator of F be available (see Section 4).

Assuming $\hat{\lambda}$ and $\hat{\underline{\beta}}(\hat{\lambda})$ are consistent,

$$f(\hat{\lambda},\underline{x}_0\hat{\underline{\beta}}(\hat{\lambda})) \to f(\lambda,\underline{x}_0\underline{\beta}) \ . \tag{2.2}$$

A Taylor expansion shows that

$$\begin{aligned} [f(\hat{\lambda},\underline{x}_0\hat{\underline{\beta}}(\hat{\lambda})) - f(\lambda,\underline{x}_0\underline{\beta})]/g(\lambda,\underline{x}_0\underline{\beta}) \\ \doteq \{\underline{x}(\hat{\underline{\beta}}(\hat{\lambda})-\underline{\beta}) + h(\lambda,\underline{x}_0\underline{\beta})(\hat{\lambda}-\lambda)\} \ , \end{aligned} \tag{2.3}$$

where

$$\begin{aligned} g(\lambda,\theta) &= f(\lambda,\theta)/(1+\lambda\theta) \\ h(\lambda,\theta) &= \theta/\lambda - \{(1+\lambda\theta)\log(1+\lambda\theta)\}/\lambda^2 \ . \end{aligned}$$

Noting that $\hat{\lambda}$ and $\underline{\hat{\beta}}(\hat{\lambda})$ are unstable and highly correlated, the expansion (2.3) shows that our problem as presently formulated is quite similar to a prediction problem in regression when there is multicollinearity.

Note: If $1 + \hat{\lambda}\underline{x}_0\underline{\hat{\beta}}(\hat{\lambda}) < 0$, then f in (2.1) must be defined by absolute values. This circumstance will not occur often; if $\hat{\lambda} > 0$ (<0), then except in rare cases $\underline{x}_0\underline{\hat{\beta}}(\hat{\lambda}) \geq \min y_i^{(\hat{\lambda})}$ ($\leq \max y_i^{(\hat{\lambda})}$), in which case $1 + \lambda\underline{x}_0\underline{\hat{\beta}}(\hat{\lambda}) \geq 0$.

Case I ($\lambda = 0$; $\sigma = 1$; normal errors; simple linear regression with slope β_1, intercept β_0)

For this special case, likelihood calculations (cf. Hinkley (1975)) can be made. Here the correct scale is the log scale and $Ec_i = 0$, $Ec_i^2 = 1$, $Ec_i^3 = \mu_3$ and $Ec_i^4 = \mu_4$. Lengthy likelihood analysis shows

$$n\,\mathrm{Cov}(\hat{\lambda},\hat{\beta}_0(\hat{\lambda}),\hat{\beta}_1(\hat{\lambda}),\hat{\sigma}(\hat{\lambda})) \to \begin{pmatrix} \Sigma_0 & B \\ B & A \end{pmatrix},$$

where

$$\Sigma_0 = 2\gamma^{-1}\begin{pmatrix} 1 & -c & \beta_0\beta_1^* \\ -c & \frac{\gamma}{2}+c^2 & -c\beta_0\beta_1^* \\ \beta_0\beta_1^* & -c\beta_0\beta_1^* & \frac{\gamma}{2}+\beta_0^2\beta_1^{*2} \end{pmatrix}$$

and where

$$c = -\tfrac{1}{2}(1+\beta_0^2+\beta_1^2)$$

$$\beta_1^* = \beta_1 + \beta_1^2\mu_3/(2\beta_0)$$

$$\gamma = 3 + 4\beta_1^2 + \beta_1^4(\mu_4-\mu_3^2-1)\ .$$

Theorem 1

$$(2.4) \qquad \frac{\mathrm{Var}[(f(\hat{\lambda},\underline{x}_0\hat{\underline{\beta}}(\hat{\lambda}))-f(\lambda=0,\underline{x}_0\underline{\beta}))\exp(-\underline{x}_0\underline{\beta})]}{\mathrm{Var}[(f(\lambda=0,\underline{x}_0\hat{\underline{\beta}})-f(\lambda=0,\underline{x}_0\underline{\beta}))\exp(-\underline{x}_0\underline{\beta})]} \to H(\beta_1) \ ,$$

where

$$H(\beta_1) = \tfrac{1}{2}(1+\beta_1^4(\mu_4-1-\mu_3^2))(6+8\beta_1^2+\beta_1^4(\mu_4-1-\mu_3^2))^{-1} + 1 \ . \qquad \square$$

The normalization by $\exp(-\underline{x}_0\underline{\beta}) = (g(\lambda=0\,,\,\underline{x}_0\underline{\beta}))^{-1}$ is as in (2.3) and is made to simplify the calculations and results. Since the numerator of (2.4) is the (suitably normalized) variance of the prediction error when λ is estimated while the denominator is the same variance for the case λ known, Theorem 1 tells us:

(i) there is a cost due to estimating λ, but it cannot exceed 50%;

(ii) for the balanced two-sample problem ($c_i = \pm 1$ with probability $\frac{1}{2}$), the cost is at most 8% and decreases to zero as $\beta_1 \to \infty$.

Case II (symmetric errors, p parameter regression, any λ)

Here we use the asymptotic theory ($n\to\infty, \sigma\to 0$) of B-D. They find that (equation (8.2)) $n^{\frac{1}{2}}(\hat{\sigma}(\hat{\lambda})-\sigma)/\sigma$ has variance $\frac{1}{2}$ and is asymptotically independent of $n^{\frac{1}{2}}(\hat{\lambda},\hat{\underline{\beta}}(\hat{\lambda}))/\sigma$, which has limiting covariance

$$\Sigma_1 = e^{-1}\begin{pmatrix} 1 & -D \\ -D' & eI+D'D \end{pmatrix} \ ,$$

where

$$\underline{x} = (1\ x_2 \ldots x_p) = (x_1 \ldots x_p)$$

$$H(a,\lambda) = \lambda^{-1}a - \lambda^{-2}(1+\lambda a)\log(1+\lambda a)$$

$$D = EH(\underline{x}\,\underline{\beta},\lambda)\underline{x}$$

$$e = E[H(\underline{x}\,\underline{\beta},\lambda)]^2 - \sum_{j=1}^{p} (Ex_j H(\underline{x}\,\underline{\beta},\lambda))^2 .$$

It is interesting to note that in the case of simple linear regression with $\lambda = 0$, Σ_1 is different from but of the same form as Σ_0 (replace c by $c_* = c + \frac{1}{2}$ and $\gamma/2$ by $e = \beta_1^4(\mu_4 - \mu_3 - 1)/4)$.

Theorem 2. In Theorem 1, replace $\lambda = 0$ by λ and $\exp(-\underline{x}_0\underline{\beta})$ by $(\sigma g(\lambda, \underline{x}_0\underline{\beta}))^{-1}$. Then Theorem 1 holds with $H(\beta_1) = 1 + 1/p$. □

The small σ asymptotics of B-D tell us that there is a positive but bounded cost due to estimating λ, with the cost decreasing as p increases. Note that Theorem 2 and Theorem 1 agree for simple linear regression, $\lambda = 0$, $\mu_4 - 1 - \mu_3^2 > 0$ and $\beta_1 \to \infty$.

B-D and Carroll also simultaneously introduced robust estimates of $(\lambda,\underline{\beta})$ based on the ideas of Huber (1977). One can use the B-D small σ asymptotics to show that (i) the cost in robust estimation for estimating λ is still 1/p and (ii) the Bickel-Doksum and Carroll methods have better robustness properties than does maximum likelihood.

We conducted a small Monte-Carlo study to check small sample performance and to investigate the results of Theorems 1 and 2. The observations were generated according to

$$(1+\beta_0+\beta_1 c_i+\varepsilon_i)^{1/\lambda} \qquad \lambda = -1,1$$

$$\exp(\beta_0+\beta_1 c_i+\varepsilon_i) \qquad \lambda = 0 \qquad (i = 1,\ldots,N) .$$

Here N=20, $\{\varepsilon_i\}$ are standard normal, $\beta_0 = 5$, $\beta_1 = 1$ and the c_i centered at zero, equally spaced, satisfy $\Sigma c_i^2 = N$ and range from -1.65 to 1.65. Then $\mu_4 = 1.79$ and $H(\beta_1) = 1.06$, so that Theorems 1 and 2 lead us to expect very little cost due to estimating λ. There were 600 iterations in the experiment. For this example,

$$\Sigma_0 = \begin{pmatrix} .27 & 3.65 & 1.35 \\ & 50.28 & 18.25 \\ & & 7.76 \end{pmatrix}$$

$$\text{Correlation Matrix} = \begin{pmatrix} 1 & .99 & .93 \\ & 1 & .92 \\ & & 1 \end{pmatrix},$$

which illustrates the multicollinearity quite well.

In Table 1 we provide an analysis of the estimates $\hat{\beta}_0(\hat{\lambda})$ and $\hat{\beta}_1(\hat{\lambda})$. The estimates are quite biased and, as noted by Carroll and B-D, the variances of $\hat{\beta}_0(\hat{\lambda}), \hat{\beta}_1(\hat{\lambda})$ are much larger than those of $\hat{\beta}_0, \hat{\beta}_1$ when λ is known. In Table 2 we present the results for the prediction problem. The first two rows are the relative biases when λ is known or unknown, respectively. In the third row we have the prediction variance when $\hat{\lambda}$ is estimated relative to that when λ is known. These tables support our asymptotic calculations and show that there is very little cost involved in estimating λ for prediction.

To this point we have defined the cost of estimating λ as an average over the distribution of the new value $\underline{x}_0$. It is also of interest to study the costs conditional on a given value of $\underline{x}_0$. For Case I when $\underline{x}_0 = (1,c_0)$, the

asymptotic ratio of the conditional prediction variances (when λ is estimated versus known) is given by

$$(2.5) \qquad \lim_{N\to\infty} \frac{\mathrm{Var}[f(\hat{\lambda},\underline{x}_0\hat{\underline{\beta}}(\hat{\lambda}))-f(\lambda=0,\underline{x}_0\underline{\beta})\,|\,\underline{x}_0=(1,c_0)]}{\mathrm{Var}[f(\lambda=0,\underline{x}_0\hat{\underline{\beta}})-f(\lambda=0,\underline{x}_0\underline{\beta})\,|\,\underline{x}_0=(1,c_0)]} = T_0(c_0,\underline{\beta}) \ ,$$

while for Case II this limit ($N\to\infty,\sigma\to 0$) is $T_1(c_0,\underline{\beta})$, where

$$T_j(c_0,\underline{\beta}) = \underline{a}\,\Sigma_j\,\underline{a}' \qquad (j = 1,2)$$

$$a = (-(\beta_0+c_0\beta_1)^2/2 \quad 1 \quad c_0) \ .$$

In Table 3 we list the values of $T_0(c_0,\underline{\beta})$, $T_1(c_0,\underline{\beta})$ and the results of a Monte-Carlo experiment (600 iterations) for the simple linear regression design discussed previously (at the different values $c_1,\ldots,c_{20}$ in this design). As expected from Theorems 1 and 2, there is only a slight cost due to estimating λ and the B-D small σ asymptotics are somewhat conservative.

Table 1

Biases and relative costs in the simple linear regression model when λ is estimated ($\beta_0=5, \beta_1=1$).

λ	$E\hat{\beta}_0(\hat{\lambda})-\beta_0$	$E\hat{\beta}_1(\hat{\lambda})-\beta_1$	$\left\{\frac{\mathrm{Var}(\hat{\beta}_0(\hat{\lambda}))}{\mathrm{Var}(\hat{\beta}_0)}\right\}^{1/2}$	$\left\{\frac{\mathrm{Var}(\hat{\beta}_1(\hat{\lambda}))}{\mathrm{Var}(\hat{\beta}_1)}\right\}^{1/2}$
1.0	.44	.20	12.9	4.0
0.0	.60	.26	9.6	4.0
-1.0	.44	.20	12.9	4.0

Table 2

Prediction biases and relative variance in the simple linear regression model.

	$\lambda = 1$	$\lambda = 0$	$\lambda = -1$
$\left\lvert \frac{E(f(\lambda,\underline{x}_0\hat{\underline{\beta}})-f(\lambda,\underline{x}_0\underline{\beta}))}{Ef(\lambda,\underline{x}_0\underline{\beta})} \right\rvert$	.00	.07	.00
$\left\lvert \frac{E(f(\hat{\lambda},\underline{x}_0\hat{\underline{\beta}}(\hat{\lambda}))-f(\lambda,\underline{x}_0\underline{\beta}))}{Ef(\lambda,\underline{x}_0\underline{\beta})} \right\rvert$	.01	.07	.00
$\frac{Var[f(\hat{\lambda},\underline{x}_0\hat{\underline{\beta}}(\hat{\lambda}))-f(\lambda,\underline{x}_0\underline{\beta})]}{Var[f(\lambda,\underline{x}_0\hat{\underline{\beta}})-f(\lambda,\underline{x}_0\underline{\beta})]}$	1.06	1.06	1.02

Table 3

Relative conditional costs in the simple linear regression model when the data are generated by $\exp(5+c+\varepsilon)$.

c_0	Likelihood analysis using Σ_0	Small σ analysis using Σ_1	Monte-Carlo
-1.65	1.01	2.00	1.08
-1.47	1.00	1.55	1.08
-1.30	1.00	1.23	1.09
-1.13	1.02	1.04	1.11
- .95	1.04	1.01	1.14
- .78	1.08	1.12	1.19
- .61	1.13	1.37	1.23
- .43	1.19	1.70	1.30
- .26	1.24	2.03	1.33
-.087	1.27	2.24	1.35
0 (not in design)	1.27	2.27	1.35
.087	1.27	2.24	1.32
.26	1.24	2.03	1.25
.43	1.19	1.70	1.18
.61	1.13	1.37	1.10
.78	1.08	1.12	1.04
.95	1.04	1.01	.99
1.13	1.02	1.04	.98
1.30	1.00	1.23	.98
1.47	1.00	1.55	1.02
1.65	1.01	2.00	1.12
Average	1.11	1.56	1.15
Predicted Average	1.06 (Theorem 1)	1.50 (Theorem 2)	—

3. Two samples: estimating the difference in medians

The balanced two-sample problem has the structure

$$y_i^{(\lambda)} = \theta_1 + \varepsilon_i \qquad i = 1,\ldots,N$$

$$= \theta_2 + \varepsilon_i \qquad i = N+1,\ldots,2N \ .$$

B-D show that for estimating $\theta_1 - \theta_2$ there is a substantial penalty for not knowing λ *unless* $\theta_1 = \theta_2$. The "treatment difference" in the original scale that we estimate is defined by

$$\Delta = f(\lambda,\theta_1) - f(\lambda,\theta_2) \ .$$

When the errors are symmetric, Δ is just the difference in the medians of the two populations. Because the B-D small σ asymptotics do not apply, we only consider the case σ fixed. We confine our attention to $\lambda = 0$ and standard normal errors. Lengthy likelihood calculations show that

$$N \ \mathrm{Cov}(\hat{\lambda},\hat{\theta}_1(\hat{\lambda}),\hat{\theta}_2(\hat{\lambda})) \to 2Q \ ,$$

where

$$Q = \gamma^{-1} \begin{pmatrix} 4 & 2\alpha & 2\beta \\ & \gamma+\alpha^2 & \alpha\beta \\ & & \alpha+\beta^2 \end{pmatrix}$$

and

$$\alpha = 1 + \theta_1^2$$

$$\beta = 1 + \theta_2^2$$

$$\mu = \theta_1 + \theta_2$$

$$\gamma = 14 + 10(\theta_1^2+\theta_2^2) + \theta_1^4 + \theta_2^4 - \alpha^2 - \beta^2 - 4\mu^2 \ .$$

Define

$$c_1 = (\theta_2^2 \exp(\theta_2) - \theta_1^2 \exp(\theta_1))/2$$

$$c_2 = \exp(\theta_1) \qquad c_3 = -\exp(\theta_2) .$$

Then, by Taylor expansions, it is possible to show

Theorem 3

$$\frac{\mathrm{Var}[f(\hat{\lambda},\hat{\theta}_1(\hat{\lambda})) - f(\hat{\lambda},\hat{\theta}_2(\hat{\lambda}))]}{\mathrm{Var}[f(\lambda=0,\hat{\theta}_1) - f(\lambda=0,\hat{\theta}_2)]} \to H(\theta_1,\theta_2) ,$$

where

$$H(\theta_1,\theta_2) = \frac{(c_1c_2c_3)Q(c_1c_2c_3)'}{c_2^2+c_3^2} .$$

□

If $\theta_1 = \theta_2$, then $H(\theta_1,\theta_2) = 1$ and there is no cost for estimating λ. Our numerical calculations indicate that $1 \le H(\theta_1,\theta_2) \le 1.03$, which indeed means there is a very small cost overall.

We conducted a Monte-Carlo experiment with $\theta_1=4,\theta_2=6$. The sample size for each population was 10 and there were 600 iterations of the experiment. Here $H(\theta_1,\theta_2) = 1.026$ and

$$Q = \begin{pmatrix} .14 & 1.21 & 2.64 \\ & 11.32 & 22.46 \\ & & 49.89 \end{pmatrix}$$

$$\text{Correlation Matrix} = \begin{pmatrix} 1 & .96 & .99 \\ & 1 & .95 \\ & & 1 \end{pmatrix} .$$

The results are given in Tables 4 and 5 from which it appears that when working in the original scale, and estimating population medians, there is essentially no cost for estimating λ.

Table 4

Biases and relative costs in the two-sample problem when λ is estimated. The populations have transformed means $\theta_1=4, \theta_2=6$, so that $\beta_0 = (\theta_1+\theta_2)/2 = 5$ and $\beta_1 = (\theta_2-\theta_1)/2 = 1$.

λ	$E\hat{\beta}_0(\hat{\lambda})-\beta_0$	$E\hat{\beta}_1(\hat{\lambda})-\beta_1$	$\left(\frac{\mathrm{Var}(\hat{\beta}_0(\hat{\lambda}))}{\mathrm{Var}(\hat{\beta}_0(\lambda))}\right)^{1/2}$	$\left(\frac{\mathrm{Var}(\hat{\beta}_1(\hat{\lambda}))}{\mathrm{Var}(\hat{\beta}_1(\lambda))}\right)^{1/2}$
1.0	.20	.13	13.2	4.2
0.0	.67	.28	16.3	5.8
-1.0	.21	.14	13.2	4.2

Table 5

Biases and relative costs in the two-sample problem. Here the treatment difference is $\Delta = E[f(\lambda,\theta_2) - f(\lambda,\theta_1)]$ and $H(\lambda,\theta_1,\theta_2) = f(\lambda,\theta_2) - f(\lambda,\theta_1)$.

	$\lambda = 1$	$\lambda = 0$	$\lambda = -1$
$\left\lvert\frac{H(\lambda,\hat{\theta}_1,\hat{\theta}_2)-\Delta}{\Delta}\right\rvert$	.01	.04	.00
$\left\lvert\frac{H(\hat{\lambda},\hat{\theta}_1(\hat{\lambda}),\hat{\theta}_2(\hat{\lambda}))-\Delta}{\Delta}\right\rvert$	.02	.02	.01
$\frac{E[H(\hat{\lambda},\hat{\theta}_1(\hat{\lambda}),\hat{\theta}_2(\hat{\lambda}))-\Delta]^2}{E[H(\lambda,\hat{\theta}_1,\hat{\theta}_2)-\Delta]^2}$	1.00	1.00	.99

4. Prediction of the conditional mean

The estimator in Section 2 is the median of the conditional distribution of y given x_0 (when F is symmetric). Our focus in this section is on estimating the conditional mean of y given x_0.

We sketch a general result which indicates that the cost of extra nuisance parameters (such as λ) is not large. We assume a regression model with (Y_i, X_i) having a joint density $g(y,x|\theta_0)$. As in normal theory regression we assume

$$g(y,x|\theta_0) = g_1(y|x,\theta_0)g_2(x) \ .$$

Letting $L_N(\theta)$ denote the log-likelihood, we make the usual assumptions:

$$E\frac{d}{d\theta}L_N(\theta_0) = 0$$

$$E\left[\frac{d}{d\theta}L_N(\theta_0)\right]\left[\frac{d}{d\theta}L_N(\theta_0)\right]' = -E\frac{d^2}{d\theta^2}L_N(\theta_0) = I(\theta_0) \tag{4.1}$$

$$N^{\frac{1}{2}}(\underline{\theta}_N - \underline{\theta}_0) \xrightarrow{L} N_q(0, I^{-1}(\theta_0)) \ ,$$

where θ_N is the maximum likelihood estimate and q is the dimension of the parameter θ_0. We are given a new value x_0 and wish to predict $E(Y|x_0)$; the natural estimate (usually only computable numerically) is

$$\hat{E(Y|x_0)} = \int y g_1(y|x_0,\theta_N)dy \ . \tag{4.2}$$

A Taylor expansion shows that

$$A_N(\theta_0,x_0) = N^{\frac{1}{2}}(E(Y|\hat{x}_0)-E(Y|x_0))$$

$$(4.3) \qquad \approx \int (y-E(y|x_0))\left[\frac{d}{d\theta}\log g_1(y|x_0,\theta_0)\right]N^{\frac{1}{2}}(\theta_N-\theta_0)g_1(y|x_0,\theta_0)\,dy$$

$$= \int (y-E(y|x_0))\left[\frac{d}{d\theta}\log g(y,x_0|\theta_0)\right]N^{\frac{1}{2}}(\theta_N-\theta_0)g_1(y|x_0,\theta_0)\,dy \ .$$

An overall measure of the accuracy of the prediction is $EA_N^2(\theta_0,x_0)$; (4.1), (4.3) and the Schwarz inequality show

$$E[A_N^2(\theta_0,x_0)|\text{sample}] \le \text{Var}(y-E(y|x_0))N^{\frac{1}{2}}(\theta_N-\theta_0)I(\theta_0)N^{\frac{1}{2}}(\underline{\theta}_N-\theta_0)'$$

$$(4.4) \qquad \xrightarrow{L} \text{Var}(y-E(y|x_0))\chi^2(q) \ ,$$

where $\chi^2(q)$ is a chi-square with q degrees of freedom. This suggests

$$(4.5) \qquad EA_N^2(\theta_0,x_0) \le q\ \text{Var}(y-E(y|x_0)) \ .$$

Equation (4.5) shows that in prediction with q parameters, the average squared prediction error is bounded and this bound increases in relative magnitude by r/q when r additional nuisance parameters are added. A similar result holds for the two-sample problem.

Example: Consider the transformation model (1.1) but take $\lambda = 1$; this means one uses the Box-Cox model when transformation is unnecessary. If there are p regression parameters, then $q = p + 1$ when $\lambda = 1$ is known and $EA_N^2(\theta_0,x_0) = \text{Var}(y-E(y|x_0))p$. When one estimates λ, (4.5) shows that $EA_N^2(\theta_0,x_0) \le \text{Var}(y-E(y|x_0))(p+2)$. Thus, the relative cost of estimating λ is 2/p (which agrees qualitatively with Theorem 2).

References

Andrews, D.F. (1971). A note on the selection of data transformations. *Biometrika 58,* 249-254.

Atkinson, A.C. (1973). Testing transformations to normality. *J. Royal Statist. Soc. Ser. B, 35,* 473-479.

Bickel, P.J. and Doksum, K.A. (1978). An Analysis of Transformations Revisited. Unpublished manuscript. University of California, Berkeley.

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

Carroll, R.J. (1978). A Robust Method for Testing Transformations to Achieve Approximate Normality. *Institute of Statistics Mimeo Series #1190.* University of North Carolina at Chapel Hill. To appear in *J. Royal Statist. Soc. Ser. B.*

Hinkley, D.V. (1975). On power transformations to symmetry. *Biometrika 62,* 101-111.

Huber, P.J. (1977). *Robust Statistical Procedures. SIAM.* Philadelphia, PA.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* On Prediction and the Power Transformation Family		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL 6. PERFORMING ORG. REPORT NUMBER Mimeo Series No. 1264
7. AUTHOR(s) R.J. Carroll and David Ruppert		8. CONTRACT OR GRANT NUMBER(s) Grant AFOSR-75-2796 Grant AFOSR-80-0080 Grant MCS78-01240
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Bolling Air Force Base Washington, D.C. 20332		12. REPORT DATE February 1980 13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for Public Release: Distribution Unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Power transformations, Prediction, Robustness, Box-Cox family, Asymptotic theory.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The power transformation family studied by Box and Cox (1964) for transforming to a normal linear model has recently been further studied by Bickel and Doksum (1978), who show that "the cost of not knowing λ and estimating it . . . is generally severe"; some of the variances for regression parameters can be extremely large. We consider prediction of future observations (untransformed) when the data can be transformed to a linear model and show that while there is a cost due to estimating λ it is generally not severe. Similar results emerge for the two sample problems. Monte-Carlo

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

20. results lend credence to the asymptotic calculations.