On prediction and the power transformation family

BY R. J. CARROLL AND DAVID RUPPERT

Department of Statistics, University of North Carolina, Chapel Hill

SUMMARY

The power transformation family is often used for transforming to a normal linear model. The variance of the regression parameter estimators can be much larger when the transformation parameter is unknown and must be estimated, compared to when the transformation parameter is known. We consider prediction of future untransformed observations when the data can be transformed to a linear model. When the transformation must be estimated, the prediction error is not much larger than when the parameter is known.

Some key words: Asymptotic distribution; Box–Cox family; Maximum likelihood estimation; Monte-Carlo simulation; Prediction of conditional median; Robustness.

1. INTRODUCTION

The power transformation family studied by Box & Cox (1964) takes the following form: for some unknown λ and $i = 1, \ldots, n$,

$$y_i^{(\lambda)} = x_i \beta + \sigma \varepsilon_i, \quad x_i = (1, c_{i2}, \ldots, c_{ip}), \quad \beta' = (\beta_0, \ldots, \beta_{p-1}). \tag{1·1}$$

Here σ is the standard deviation; the ε_i are independently and identically distributed with mean zero, variance one and distribution F, and

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & (\lambda \neq 0), \\ \log y & (\lambda = 0). \end{cases}$$

Box & Cox propose maximum likelihood estimates for λ and β when F is the normal distribution. There are numerous alternative methods as well as proposals for testing hypotheses of the form H_0: $\lambda = \lambda_0$ (Hinkley, 1975; Andrews, 1971; Atkinson, 1973; Carroll, 1980). Carroll studied the testing problem via Monte-Carlo; by allowing F to be nonnormal he approximated a problem with outliers and found that the chance of mistakenly rejecting the null hypothesis can be very high indeed.

Bickel & Doksum (1981) develop an asymptotic theory for estimation. For technical reasons they assume that the design vectors $x_1, x_2, \ldots$ are independent and identically distributed according to G. If the maximum likelihood estimate of the regression parameter is $\hat\beta$ when λ is known, and $\beta^* = \hat\beta(\hat\lambda)$ when λ is unknown and estimated by $\hat\lambda$, they compute the asymptotic distributions of $n^{\frac{1}{2}}(\hat\beta - \beta)/\sigma$ and $n^{\frac{1}{2}}(\beta^* - \beta)/\sigma$ as $n \to \infty$ and $\sigma \to 0$. These distributions, which are given in the Appendix, are different, and as regards variances

> the cost of not knowing λ and estimating it . . . is generally severe. . . . The problem is that β^* and $\hat\lambda$ are highly correlated.

Their theoretical and Monte Carlo work indicate that $\hat{\lambda}$ and β^* are highly variable and highly correlated, and as discussed in § 2, the problem is similar in nature to that of multicollinearity. An example of the variability of β^* is given in the next section.

These results are somewhat controversial. 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, as in Bickel & Doksum's 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 x arrive independently with the distribution G. Motivated by this fact, we focus our attention specifically on prediction, but we also discuss the two-sample problem and a somewhat more general estimation theory. Using Bickel & Doksum's asymptotic theory and Monte Carlo, we find that for prediction as well as 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

2·1. *The general case*

Our model specifically includes an intercept, i.e. $x_i = (1, c_i)$; by suitable rescaling we assume the c_i have mean zero and identity covariance. From the sample we calculate $\hat{\lambda}$ and β^*, and we are given a new vector $x_0 = (1, c_0)$, which is independent of the other x's but still has the same distribution G. This formulation is simple but hardly necessary; the design vectors x_i could satisfy the usual regression assumptions, and x_0 can be thought of as chosen according to the design. Our predicted value in the transformed scale would be $x_0\beta^*$, so a natural predictor is $f(\hat{\lambda}, x_0\beta^*)$ where

$$f(\lambda, \theta) = \begin{cases} (1+\lambda\theta)^{1/\lambda} & (\lambda \neq 0), \\ e^{\theta} & (\lambda = 0). \end{cases}$$

Notice that if F has median equal to 0, then $f(\lambda, x_0\beta)$ 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 § 3 for further discussion.

A Taylor expansion shows that

$$f(\hat{\lambda}, x_0\beta^*) - f(\lambda, x_0\beta)/g(\lambda, x_0\beta) \simeq x_0(\beta^* - \beta) + h(\lambda, x_0\beta)(\hat{\lambda} - \lambda) \qquad (2{\cdot}1)$$

where

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

Estimates $\hat{\lambda}$ and β^* are unstable and highly correlated, and expansion (2·1) shows that our problem as presently formulated is quite similar to a prediction problem in regression when there is multicollinearity.

2·2. *Case* 1

We now assume that F is a normal distribution, $\lambda = 0$, $\sigma = 1$, and the model is simple linear regression with slope β_1 and intercept β_0.

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

$$n \operatorname{cov}(\hat{\lambda}, \beta_0^*, \beta_1^*) \rightarrow \Sigma_0,$$

where

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

and where

$$c = -\tfrac{1}{2}(1+\beta_0^2+\beta_1^2), \quad \beta_1^* = \beta_1 + \tfrac{1}{2}\beta_1^2 \mu_3/\beta_0, \quad \gamma = 3+4\beta_1^2+\beta_1^4(\mu_4-\mu_3^2-1).$$

Note that if λ were not estimated we would have had Σ_0 as the identity matrix, and in the next section we give an example which demonstrates the multicollinearity.

THEOREM 1. *Let* MSE (λ, x_0) *be the mean squared error for estimating the conditional median of Y given x_0 and λ known, while* MSE $(\hat{\lambda}, x_0)$ *is the same quantity but with λ unknown. Then*

$$E_G\left(\| x_0 \|^2 \frac{\text{MSE}\,(\hat{\lambda}, x_0)}{\text{MSE}\,(\lambda, x_0)} \right) \Big/ E(\| x_0 \|^2) \rightarrow H(\beta_1), \tag{2·2}$$

where

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

Note that $\mu_4 - 1 - \mu_3^2 = E\{(c_1^2 - \mu_3 c_1 - 1)^2\} \geqslant 0$. The quantity (2·2) is a modified form of the average cost for prediction when λ is estimated. If one prefers to assume the design vectors are constants, then one might think of (2·2) as an average over the design. In either case the results are encouraging:

(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 $\tfrac{1}{2}$, the cost is at most 8% and decreases to zero as $\beta_1 \rightarrow \infty$.

2·3. *Case 2: Symmetric errors*

We now allow λ and the number of regression parameters, p, to be arbitrary, but we assume that F is symmetric about zero.

Here we use the asymptotic theory of Bickel & Doksum, in which $n \rightarrow \infty$ and $\sigma \rightarrow 0$ simultaneously; see the Appendix for details. We report results only for the simplest case of an orthogonal design in which

$$n^{-1} \sum_{i=1}^{n} x_i' x_i \rightarrow I.$$

It then follows that $(\lambda, \hat{\beta}^*)$ is asymptotically normally distributed with mean (λ, β) and covariance $\sigma\Sigma_1/n$, where

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

and

$$x = (1, x_2, \ldots, x_p) = (x_1, \ldots, x_p), \quad H(a, \lambda) = \lambda^{-1} a - \lambda^{-2}(1+\lambda a)\log(1+\lambda a),$$

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

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

THEOREM 2. *As $N \to \infty$ and $\sigma \to 0$ for any λ,*

$$E_G\left\{\|x_0\|^2 \frac{\text{MSE}(\hat{\lambda}, x_0)}{\text{MSE}(\lambda, x_0)}\right\} \Big/ E_G(\|x_0\|^2) \to 1 + 1/p,$$

where p is the dimension of the vector β.

The small σ asymptotics of Bickel & Doksum 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 $\beta_1 \to \infty$.

Bickel & Doksum and Carroll also simultaneously introduced robust estimates of (λ, β) based on the ideas of Huber (1977). One can use Bickel & Doksum's small σ asymptotics to show that (i) the cost in robust estimation for estimating λ is still $1/p$ and (ii) Bickel & Doksum's and Carroll's 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 $(1 + \beta_0 + \beta_1 c_i + \varepsilon_i)^{1/\lambda}$ for $\lambda = -1$, and $\exp(\beta_0 + \beta_1 c_i + \varepsilon_i)$ for $\lambda = 0$. Here $n = 20$, the ε_i are standard normal, $\beta_0 = 5$, $\beta_1 = 1$ and the c_i centred at zero, equally spaced, satisfy $\Sigma c_i^2 = n$ and range from $-1{\cdot}65$ to $1{\cdot}65$. Then $\mu_4 = 1{\cdot}79$ and $H(\beta_1) = 1{\cdot}06$, so that Theorems 1 and 2 lead us to expect very little cost due to estimating λ. There were 600 repetitions of the experiment. Likelihood calculations show that

$$\Sigma_0 = \begin{bmatrix} 0{\cdot}27 & 3{\cdot}65 & 1{\cdot}35 \\ \cdot & 50{\cdot}28 & 18{\cdot}25 \\ \cdot & \cdot & 7{\cdot}76 \end{bmatrix}$$

with correlation matrix

$$\begin{bmatrix} 1 & 0{\cdot}99 & 0{\cdot}93 \\ \cdot & 1 & 0{\cdot}92 \\ \cdot & \cdot & 1 \end{bmatrix},$$

which illustrates the multicollinearity quite well, for if λ were known then $n^{\frac{1}{2}}(\hat{\beta}_0 - \beta_0)$ and $n^{\frac{1}{2}}(\hat{\beta}_1 - \beta_0)$ would be uncorrelated with common variance 1.

In rows 1 to 4 of Table 1, we provide an analysis of the estimates β_0^* and β_1^* in the case that λ is estimated. The estimates are biased and have much larger mean squared errors than the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained for the case that λ is known.

The remaining rows of Table 1 give the results for the prediction problem. The last row corresponds to Theorems 1 and 2, although the actual mean squared errors are computed. It appears that, on the average, our asymptotic calculations are reasonable, and there is

Table 1. *Monte-Carlo results for the model* $y_i = \beta_0 + \beta_1 c_i + \sigma\epsilon_i$, $\beta_0 = 5$, *and* $\beta_1 = 1$; E_K *and* E_U *denote expectation when* λ *is known and unknown, respectively*

	$\lambda = -1{\cdot}0$	$\lambda = 0{\cdot}0$
$\lvert E_U(\hat{\beta}_0) - \beta_0 \rvert$	0·44	0·60
$\lvert E_U(\hat{\beta}_1) - \beta_1 \rvert$	0·20	0·26
$[E_U\{(\hat{\beta}_0 - \beta_0)^2\}/E_K\{(\hat{\beta}_0 - \beta_0)^2\}]^{\frac{1}{2}}$	12·9	9·6
$[E_U\{(\hat{\beta}_1 - \beta_1)^2\}/E_K\{(\hat{\beta}_1 - \beta_1)^2\}]^{\frac{1}{2}}$	4·0	4·0
$\dfrac{E_U[\{f(\hat{\lambda}, \hat{\beta}_0) - f(\lambda, \beta_0)\}^2]}{E_K[\{f(\lambda, \hat{\hat{\beta}}_0) - f(\lambda, \beta_0)\}^2]}$	—	1·35 1·27* 2·27†
$\dfrac{E_U[\{f(\hat{\lambda}, \hat{\beta}_0 - 1{\cdot}65\hat{\beta}_1) - f(\lambda, \beta_0 - 1{\cdot}65\beta_1)\}^2]}{E_K[\{f(\lambda, \hat{\hat{\beta}}_0 - 1{\cdot}65\hat{\hat{\beta}}_1) - f(\lambda, \beta_0 - 1{\cdot}65\beta_1)\}^2]}$	—	1·08 1·01* 2·00†
$\dfrac{E_U[\{f(\hat{\lambda}, \hat{\beta}_0 + \hat{\beta}_1 c_0) - f(\lambda, \beta_0 + \beta_1 c_0)\}^2]}{E_K[\{f(\lambda, \hat{\hat{\beta}}_0 + \hat{\hat{\beta}}_1 c_0) - f(\lambda, \beta_0 + \beta_1 c_0)\}^2]}$	1·02	1·06

* The value predicted by a likelihood analysis using Σ_0.

† The value predicted by the small σ analysis using Σ_1.

For the last entry, c_0 is randomly chosen from the design.

only a small cost involved in estimating λ for prediction. To read rows 5 and 6, we note that to this point we have defined the cost of estimating λ as an average over the distribution of the new value x_0. It is also of interest to study the costs conditional on a given value of x_0. For Case 1 when $x_0 = (1, c_0)$ and $\lambda = 0$ we find that

$$\frac{\text{MSE}\,(\hat{\lambda}, x_0)}{\text{MSE}\,(\lambda = 0, x_0)} \rightarrow \Upsilon_0(c_0, \beta),$$

while for Case 2 this limit is $\Upsilon_1(c_0, \beta)$, where

$$\Upsilon_j(c_0, \beta) = a\,\Sigma_j\,a^{\mathrm{T}} \quad (j = 1, 2), \quad a = [-\tfrac{1}{2}(\beta_0 + \beta_1 c_0)^2, 1, c_0).$$

Rows 5 and 6 of Table 1 give the ratios of the mean squared errors at two points, the centre and an extreme of the design. As expected from Theorems 1 and 2, there is only a slight cost due to estimating λ, and the small σ asymptotics of Bickel & Doksum are somewhat conservative.

3. Prediction of the conditional mean

The estimator in § 2 is the median of the conditional distribution of y given x_0. 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 $g(y, x \mid \theta_0)$. As in normal theory regression we assume

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

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

$$\begin{gathered} E\{L_n'(\theta_0)\} = 0, \\ E\{L_n'(\theta_0)\, L_n'(\theta_0)^{\mathrm{T}}\} = -E\{L_n''(\theta_0)\} = I(\theta_0), \\ n^{\frac{1}{2}}(\theta_n - \theta_0) \rightarrow N_q\{0, I^{-1}(\theta_0)\}, \end{gathered} \tag{3·1}$$

where θ_n is the maximum likelihood estimate, q is the dimension of the parameter θ_0 and the prime denotes differentiation with respect to θ at $\theta = \theta_0$. We are given a new value x_0 and wish to predict $E(Y\,|\,x_0)$; the natural estimate, which usually is only computable numerically, is

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

Taylor expansion shows that

$$\begin{aligned} A_n(\theta_0, x_0) &= n^{\frac{1}{2}}\{(\hat{E}(Y\,|\,x_0) - E(Y\,|\,x_0)\} \\ &\simeq \int \{y - E(y\,|\,x_0)\}\left\{\frac{d}{d\theta}\log g_1(y\,|\,x_0, \theta_0)\right\} n^{\frac{1}{2}}(\theta_n - \theta_0)\, g_1(y\,|\,x_0, \theta_0)\,dy \\ &= \int \{y - E(y\,|\,x_0)\}\left[\frac{d}{d\theta}\log g(y, x_0\,|\,\theta_0)\right] n^{\frac{1}{2}}(\theta_n - \theta_0)\, g_1(y\,|\,x_0, \theta_0)\,dy. \qquad (3{\cdot}2) \end{aligned}$$

An overall measure of the accuracy of the prediction is $E\{A_n^2(\theta_0, x_0)\}$; (3·1) and (3·2) and Schwarz's inequality show that for a sample $\mathscr{S}$

$$E\{A_n^2(\theta_0, x_0)\,|\,\mathscr{S}\} \leqslant \operatorname{var}\{y - E(y\,|\,x_0)\}\, n^{\frac{1}{2}}(\theta_n - \theta_0)^{\mathrm{T}} I(\theta_0)\, n^{\frac{1}{2}}(\theta_n - \theta_0).$$

Since $n^{\frac{1}{2}}(\theta_n - \theta_0)^{\mathrm{T}} I(\theta_0)\, n^{\frac{1}{2}}(\theta_n - \theta_0)$ converges in distribution to a chi-squared variable with q degrees of freedom, this suggests that

$$E\{A_n^2(\theta_0, x_0)\} \leqslant q \operatorname{var}\{y - E(y\,|\,x_0)\}. \qquad (3{\cdot}3)$$

Equation (3·3) 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

$$E\{A_n^2(\theta_0, x_0)\} = \operatorname{var}\{y - E(y\,|\,x_0)\}\, p.$$

When one estimates λ, (3·3) shows that

$$E\{A_n^2(\theta_0, x_0)\} \leqslant \operatorname{var}\{y - E(y\,|\,x_0)\}\,(p+2).$$

Thus, the relative cost of estimating λ is at most $2/p$, which agrees qualitatively with Theorem 2.

We thank Professors Bickel and Doksum for providing a copy of their paper and the referee for his helpful comments.

APPENDIX

Some asymptotics

Suppose that the distribution function F is symmetric. In the theory of Bickel & Doksum (1981), it is assumed that $\sigma = r\eta$ where $r = r(n)$ is a known sequence tending to zero and η is unknown and fixed. Define

$$A = (x_1, \ldots, x_n)^{\mathrm{T}}, \quad P = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}, \quad Q = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}d^{\mathrm{T}}, \quad d = (d_1, \ldots, d_n).$$

$$d_i = \{\lambda^{-2}(v_i - 1) - v_i \log|v_i|\}, \quad v_i = 1 + \lambda x_i \beta, \quad e = dd^{\mathrm{T}} - dPd^{\mathrm{T}}.$$

Assuming that e converges to a positive limit, they prove after very detailed calculations that $n^{\frac{1}{2}}\{(\hat{\lambda}-\lambda)/\sigma, (\beta^*-\beta)/\sigma, (\hat{\eta}-\eta)/\eta\}$ is asymptotically normally distributed with mean zero and covariance

$$\lim_{n\to\infty} e^{-1}\begin{bmatrix} 1 & -Q & 0 \\ -Q^{\mathrm{T}} & (n^{-1}A^{\mathrm{T}}A)^{-1}e+QQ^{\mathrm{T}} & 0 \\ 0 & 0 & \frac{1}{2}e \end{bmatrix}.$$

Hence when λ is estimated one adds to the covariance of β^* the term $\lim\,(QQ^{\mathrm{T}}e^{-1})$, which is positive-semidefinite and, as the example shows, can often be much larger than the covariance of $\hat{\beta}$ when λ is known. It is this extra term which causes the instability of the regression estimate β^* when λ is estimated.

References

Andrews, D. F. (1971). A note on the selection of data transformations. *Biometrika* **58**, 249–54.

Atkinson, A. C. (1973). Testing transformations to normality. *J. R. Statist. Soc.* B **35**, 473–9.

Bickel, P. J. & Doksum, K. A. (1981). An analysis of transformations revisited. *J. Am. Statist. Assoc.* **76**, 296–311.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc.* B **26**, 211–52.

Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. *J. R. Statist. Soc.* B **42**, 71–8.

Hinkley, D. V. (1975). On power transformations to symmetry. *Biometrika* **62**, 101–11.

Huber, P. J. (1977). *Robust Statistical Procedures*. Philadelphia: Society of Industrial and Applied Mathematics.

[*Received February* 1980. *Revised May* 1981]