A Note on Adaptive Box-Cox Transformation Parameter in Linear Regression

Mezbahur Rahman, Samah Al-thubaiti, Reid W Breitenfeldt, Jinzhu Jiang, Elliott B Light, Rebekka H Molenaar, Mohammad Shaha A Patwary, and Joshua A Wuollet
Correspondence email: mezbahur.rahman@mnsu.edu

Minnesota State University
Mankato, MN 56001, USA

Abstract

The Box-Cox transformation is a well known family of power transformations that brings a set of data into agreement with the normality assumption of the residuals and hence the response variable of a postulated model in regression analysis. In this paper we use six different data sets to implement adaptive maximum likelihood Box-Cox transformation parameter estimation in regression analysis. In addition, we perform random permutation and Monte-Carlo simulation to investigate the performances of the adaptive method.

Key Words: Moments for the ordered standard normal variates; Normality tests; Shapiro-Wilk W statistic.

1. Introduction

In regression analysis, often the key assumption regarding normality of the error variable and hence the response variable are violated. The commonly used remedy is the Box-Cox family of power transformations (Box and Cox (1964)). The process is to select a parameter in the Box-Cox transformation which maximizes the normal likelihood using the data at hand and then apply regression analysis on the transformed response variable. In practice, the regression model parameters are usually estimated separately after the necessary Box-Cox power transformation parameter is selected.

In literature, the estimation procedures of the Box-Cox power transformation parameter are considered by many authors. The notable ones are the normal

likelihood method of Box and Cox (1964), the robustified version of the normal likelihood method of Carroll (1980) and of Bickel and Doksum (1981), the transformation to symmetry method of Hinkley (1975), the quick estimate of Hinkley (1977) and of Taylor (1985). Lin and Vonesh (1989) constructed a nonlinear regression model which is used to estimate the transformation parameter such that the normal probability plot of the data on the transformed scale is as close to linearity as possible. Following the footsteps of Box and Cox (1982) and Lin and Vonesh (1989), Halawa (1996) considered the power transformation parameter estimation procedure using an artificial regression model which gives the estimates with very small variabilities compared to the normal likelihood procedure. Halawa (1996) conducted an exhaustive comparative study with normal likelihood procedure. In that study, he also considered estimation procedures of the location and the scale parameters in the likelihood.

Rahman (1999) introduced a method of estimating the Box-Cox power transformation parameter using maximization of the Shapiro-Wilk W (Shapiro and Wilk (1965)) statistic along with a comparison study of the normal likelihood method (Carroll (1980)), and of the artificial regression model method (Halawa (1996)). Rahman and Pearson (2008) showed that adaptive procedure's performance is better than other alternatives in obtaining maximum likelihood estimation procedures. In this paper the estimation procedure for the Box-Cox power transformation parameter is considered using adaptive maximization of the normal likelihood along with the Newton-Raphson root finding method explicitly in multiple regression context.

The motivation of the paper is given in the following section. Shapiro-Wilk W statistic p-values are computed before and after implementation of the transformation parameter. used and hence a section is Comparisons are done using random permutation and for normal and non-normal samples using Monte-Carlo simulation.

2. Motivation

From the authors experience with computing the maximum likelihood estimate of the Box-Cox parameter in several occassions it is found that the likelihood is an monotonously increasing or decreasing function in the usual range of -2 to +2. By adjusting the range without limitation it is found that always there is a maximum. Hence we initiated to pursue Newton-Raphson method to obtain the maximum. While using the 'usual' Newton-Raphson iteration it is noticed that the estimate remains closer to the initial value and often the maximum is not achieved. Hence an adaptive method is obtained which correctly identifies the maximum. In literature, most simulations are done by just normalizing a random data set. Here we explicitly implement the procedure within multiple regression setting. We use random permutation to deviate from the original data structure. And we use Monte-Carlo simulation to consider normal and non-normal data.

3. Box-Cox Transformation

Let $Y_1, Y_2, \cdots, Y_n$ be a random sample of size n from a population whose functional form is unknown. Box and Cox (1964) suggested that if the transformation

$$Y^* = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(Y), & \lambda = 0 \end{cases} \tag{1}$$

is performed on the data then Y^* will have an approximate normal distribution with mean μ and variance σ^2. In equation (1), λ is unknown and considered as the Box-Cox power transformation parameter and 'ln' represents the natural logarithm.

4. Adaptive Newton-Raphson Method in Multiple Regression

Let us consider $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ be an usual multiple regression model, where $\mathbf{Y}$ is an $n \times 1$ vector of a dependent variable measurements, $\mathbf{X}$ is an $n \times p$ matrix of independent variable measurements, β is an $n \times 1$ vector of unknown regression

parameters, and ϵ is an $n \times 1$ vector of error variable.

The least square estimate of β is $\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$. Residuals are computed as $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. If $\mathbf{e}$ does not follow normal distribution, Box-Cox transformation on $\mathbf{Y}$ is implemented as in (1). After applying the transformation mentioned in equation (1), the density function of e in terms of y can be written as

$$f(y;\lambda,\mu,\sigma^2) \doteq \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left(\frac{y^\lambda - 1}{\lambda} - \mathbf{x'}\beta\right)^2} \cdot y^{\lambda-1}, & \lambda \neq 0, \\ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(\ln(y) - \mathbf{x'}\beta)^2} \cdot \frac{1}{y}, & \lambda = 0, \end{cases} \tag{2}$$

where $\mathbf{x'}$ is the respective row of the $\mathbf{X}$ matrix for a given y.

For a fixed λ, the log-likelihood function $\ell_A = \ell(\mathbf{x'}\beta, \sigma^2; y_1, y_2, \cdots, y_n)$, where the subscript "A" stands for adaptive, can be written as

$$\ell_A = \begin{cases} -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(\frac{y_i^\lambda - 1}{\lambda} - \mathbf{x'}\beta\right)^2 + (\lambda - 1)\sum_{i=1}^n \ln(y_i), & \lambda \neq 0, \\ -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(\ln(y_i) - \mathbf{x'}\beta\right)^2 - \sum_{i=1}^n \ln(y_i), & \lambda = 0. \end{cases} \tag{9}$$

Equation (9) is maximized when the partial derivatives of (9) with respect to β and σ^2 are equated to zero and the solution to the corresponding system is found. This leads to solving the following two equations,

$$\frac{\partial \ell_A}{\partial \beta}\Big|_{\beta=\hat{\beta}_A, \sigma^2=\hat{\sigma^2}_A} = 0 \Longrightarrow \hat{\beta}_A = (\mathbf{X'X})^{-1}\mathbf{X'Y}^* \tag{10}$$

and

$$\frac{\partial \ell_A}{\partial \sigma^2}\Big|_{\beta=\hat{\beta}_A, \sigma^2=\hat{\sigma^2}_A} = 0 \Longrightarrow \hat{\sigma^2}_A = \frac{1}{n}\left(\mathbf{Y}^* - \mathbf{X}\hat{\beta}_\mathbf{A}\right)'\left(\mathbf{Y}^* - \mathbf{X}\hat{\beta}_\mathbf{A}\right). \tag{11}$$

Then the pseudo log-likelihood $\ell_A^* = \ell(\lambda; y_1, y_2, \cdots, y_n)$ can be written as

$$\ell_A^* = -\frac{n}{2}\left[\ln(2\pi) + \ln\left\{\frac{1}{n}\sum_{i=1}^n \left(\frac{y_i^\lambda - 1}{\lambda} - \mathbf{x_i}'\hat{\beta}_A\right)^2\right\} + 1\right] + (\lambda - 1)\sum_{i=1}^n \ln y_i. \tag{12}$$

The steps for maximizing ℓ_A^* using the Newton-Raphson method are as follows:

$$\frac{\partial \ell_A^*}{\partial \lambda}\Big|_{\lambda=\hat{\lambda}_A} = 0 \Longrightarrow h(\hat{\lambda}_A) = -n\frac{\sum_{i=1}^n (P_i - U_i)(Q_i - V_i)}{\sum_{i=1}^n (P_i - U_i)^2} + \sum_{i=1}^n \ln y_i = 0 \tag{13},$$

where $P_i = y_i^* = \left(\frac{y_i^{\hat{\lambda}_A}-1}{\hat{\lambda}_A}\right)$, $Q_i = \frac{\partial y_i^*}{\partial \lambda} = \left(\frac{\hat{\lambda}_A(\ln y_i) y_i^{\hat{\lambda}_A} - y_i^{\hat{\lambda}_A}+1}{\hat{\lambda}_A^2}\right)$, $U_i = \hat{y_i}^* = \mathbf{x_i}'\hat{\beta}_A = \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'P_i$, and

$$\hat{\lambda}_A^{(t+1)} = \hat{\lambda}_A^{(t)} - \frac{h(\hat{\lambda}_A^{(t)})}{h'(\hat{\lambda}_A^{(t)})}, \tag{14}$$

where

$$h'(\hat{\lambda}_A) = -n\left[\frac{\sum_{i=1}^n (P_i-U_i)(R_i-W_i) + \sum_{i=1}^n (Q_i-V_i)^2}{\sum_{i=1}^n (P_i-U_i)^2}\right.$$

$$\left. -2\frac{\{\sum_{i=1}^n (P_i-U_i)(Q_i-V_i)\}^2}{\left\{\sum_{i=1}^n (P_i-U_i)^2\right\}^2}\right],$$

where $R_i = \frac{\partial^2 y_i^*}{\partial \lambda^2} = \left(\frac{\hat{\lambda}_A^2(\ln y_i)^2 y_i^{\hat{\lambda}_A} - 2\hat{\lambda}_A(\ln y_i) y_i^{\hat{\lambda}_A} + 2y_i^{\hat{\lambda}_A} - 2}{\hat{\lambda}_A^3}\right)$, $V_i = \frac{\partial \hat{y_i}^*}{\partial \lambda} = \frac{\partial \mathbf{x_i}'\hat{\beta}_A}{\partial \lambda} = \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Q_i$, and $W_i = \frac{\partial^2 \hat{y_i}^*}{\partial \lambda^2} = \frac{\partial^2 \mathbf{x_i}'\hat{\beta}_A}{\partial \lambda^2} = \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'R_i$.

The computation algorithm starts with an initial value of $\hat{\lambda}_A$ (an obvious choice is 1) iteratively using (14) and then $\hat{\mu}_A$ and $\hat{\sigma^2}_A$ are obtained using (10) and (11) by substituting $\hat{\lambda}_A$ for λ, if desired.

5. Shapiro-Wilk W Statistic

The Shapiro-Wilk W test statistic (Shapiro and Wilk (1965)) is obtained by dividing the square of an appropriate linear combination of the sample order statistics by the usual symmetric estimate of the variance.

Let $(X_1, X_2, \cdots, X_n)$ be a random sample to be tested for normality, ordered $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. Define

$$W = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

where the vector $\mathbf{a} = \frac{\mathbf{m}'V^{-1}}{(\mathbf{m}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{1/2}}$, $\mathbf{m}$ is the vector of the expected values and $\mathbf{V}$ is the variance covariance matrix of the standard normal order statistics.

The value of W is closer to 1 means that the data is closer to normality and the maximum value of W is 1. The Shapiro and Wilk (1965) W statistic

has been shown to yield a powerful test of normality for a variety of nonnormal distributions (Pearson, D'Agostino, and Bowman (1977) and Shapiro, Wilk, and Chen (1968)).

The values of the a_i's are tabulated in Shapiro and Wilk (1965) for $n =$ 2(1)50. For other sample sizes, the a_i's can be estimated using the following suggested approximations:

$$\hat{a}_i^* = 2m_i, \quad i = 2, 3, \cdots, n-1, \quad \text{and}$$

$$\hat{a}_1^2 = \hat{a}_n^2 = \begin{cases} \frac{\Gamma\left(\frac{1}{2}n\right)}{\sqrt{2}\Gamma\left(\frac{1}{2}(n+1)\right)}, & n \leq 20, \\ \frac{\Gamma\left(\frac{1}{2}(n+1)\right)}{\sqrt{2}\Gamma\left(\frac{1}{2}n+1\right)}, & n > 20, \end{cases}$$

then a_i^* for $i = 2, 3, \ldots, n-1$ are normalized by dividing by $C = \sqrt{-2.722 + 4.0832n}$ as suggested by Shapiro and Wilk (1965).

The values of the m_i's are tabulated in Harter (1961) for $n =$ 2(1)100, 125(25)250, 300, 350, and 400. More accurate values of the $\mathbf{m}_i$'s and $\mathbf{V}$ are also given in Parish (1992a and 1992b) for n= 2(1)50. In this study, we will use explicit Monte-Carlo simulation in finding the coefficients and hence the p-values of the Shapiro-Wilk W' statistic as suggested by Rahman and Pearson (2000).

6. Application

6.1. Data Set 1:

Table 1: Data Set 1:

Model	Prior Transformation		Post Transformation		
	W-statistic	p-value	λ-value	W-statistic	p-value
Model 1	0.9882	0.9033	2.1962	0.9691	0.2512
Model 2	0.9764	0.4746	2.0293	0.9522	0.0898

Table 3: Jet Turbine Engine Thrust Data (Table B.13, Montgomery et al., 2012, pp. 566)

y	x_1	x_2	x_3	x_4	x_5	x_6	y	x_1	x_2	x_3	x_4	x_5	x_6
4540	2140	20640	30250	205	1732	99	4330	2062	20500	30190	193	1748	101
4315	2016	20280	30010	195	1697	100	4119	1929	20050	29960	183	1713	100
4095	1905	19860	29780	184	1662	97	3891	1815	19680	29770	173	1684	100
3650	1675	18980	29330	164	1598	97	3467	1595	18890	29360	153	1624	99
3200	1474	18100	28960	144	1541	97	3045	1400	17870	28960	134	1569	100
4833	2239	20740	30083	216	1709	87	4411	2047	20540	30160	193	1746	99
4617	2120	20305	29831	206	1669	87	4203	1935	20160	29940	184	1714	99
4340	1990	19961	29604	196	1640	87	3968	1807	19750	29760	173	1679	99
3820	1702	18916	29088	171	1572	85	3531	1591	18890	29350	153	1621	99
3368	1487	18012	28675	149	1522	85	3074	1388	17870	28910	133	1561	99
4445	2107	20520	30120	195	1740	101	4350	2071	20460	30180	198	1729	102
4188	1973	20130	29920	190	1711	100	4128	1944	20010	29940	186	1692	101
3981	1864	19780	29720	180	1682	100	3940	1831	19640	29750	178	1667	101
3622	1674	19020	29370	161	1630	100	3480	1612	18710	29360	156	1609	101
3125	1440	18030	28940	139	1572	101	3064	1410	17780	28900	136	1552	101
4560	2165	20680	30160	208	1704	98	4402	2066	20520	30170	197	1758	100
4340	2048	20340	29960	199	1679	96	4180	1954	20150	29950	188	1729	99
4115	1916	19860	29710	187	1642	94	3973	1835	19750	29740	178	1690	99
3630	1658	18950	29250	164	1576	94	3530	1616	18850	29320	156	1616	99
3210	1489	18700	28890	145	1528	94	3080	1407	17910	28910	137	1569	100

where, y: Thrust (mechanical force generated by the engines) (lbs), x_1: Primary speed of rotation (rpm), x_2: Secondary speed of rotation (rpm), x_3: Fuel flow rate (lbs/min), x_4: Pressure (lbs/in^2), x_5: Exhaust temperature ($^\circ F$), and x_6: Ambient temperature at time of test ($^\circ F$)

RR Reject prior transformation and reject after transformation
RA Reject prior transformation and accept after transformation
AR Accept prior transformation and reject after transformation
AA Accept prior transformation and accept after transformation
NI Number of p-values increased after transformation

7. Concluding Remarks

In practice, $\hat{\lambda}_A$ and $\hat{\lambda}_G$ should be preferred over $\hat{\lambda}_I$. In maximizing the likelihhod, no limitation on the range of λ values should be imposed. The adap-

Table 2: Data Set 1:

Model	Simulation Type	RR	RA	AR	AA	NI
Model 1	Random Permutation	61	109	4	826	831
Model 2	Random Permutation	53	101	2	844	868
Model 1	Uniform (0,1)	0	179	0	821	576
Model 2	Uniform (0,1)	0	219	0	781	431
Model 1	Normal (0,1)	0	109	0	891	349
Model 2	Normal (0,1)	0	108	0	892	206
Model 1	Exponential (1)	0	910	0	90	973
Model 2	Exponential (1)	0	943	0	57	973
Model 1	$\frac{3}{4}N(0,1)+\frac{1}{4}N(\frac{3}{2},\frac{1}{3})$	0	172	0	828	469
Model 2	$\frac{3}{4}N(0,1)+\frac{1}{4}N(\frac{3}{2},\frac{1}{3})$	0	188	0	812	345

tive method is simple enough to implement in any computational environment. While the grid search method may readily impose a bound on the range of λ this may fail to maximize the likelihood.

11. Bibliography

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

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

Box, G. E. P. and D. R. Cox (1982). "An Analysis of Transformations Revisited (Rebutted)." *Journal of the American Statistical Association,* 77, 209-210.

Carroll, R. J. (1980). "A Robust Method for Testing Transformations to Achieve Approximate Normality." *Journal of the Royal Statistical Society, Series B.,* 42, 71-78.

Conover, W. J. (1980). *Practical Nonparametric Statistics, 2nd ed.* John Wiley & Sons, New York.

Halawa, A. M. (1996). "Estimating the Box-Cox Transformation via an Artifi-

cial Regression Model." *Communications in Statistics — Simulation and Computation,* 25(2), 331-350.

Harter, H. L. (1961). "Expected Values of Normal Order Statistics." *Biometrika,* 48, 1 and 2, 151-165.

Hinkley, D. V. (1975). "On Power Transformation to Symmetry." *Biometrika,* 62, 101-111.

Hinkley, D. V. (1977). "On Quick Choice of Power Transformation." *Applied Statistics,* 26, 67-68.

Lin, L. I. and E. F. Vonesh (1989). "An Empirical Nonlinear Data-Fitting Approach for Transforming Data to Normality." *American Statistician,* 43, 237-243.

Parish, R. S. (1992a). "Computing Expected Values of Normal Order Statistics." *Communications in Statistics - Simulation and Computation,* 21(1), 57-70.

Parish, R. S. (1992b). "Computing Variances and Covariances of Normal Order Statistics." *Communications in Statistics - Simulation and Computation,* 21(1), 71-101.

Pearson, E. S., R. B. D'Agostino, and K. O. Bowman (1977). "Tests for Departure from Normality: Comparison of Powers." *Biometrika,* 64, 231-246.

Rahman, M. and L. M. Pearson (2008). "A Note on the Maximum Likelihood Box-Cox Transformation Parameter." *Journal of Probability and Statistical Science,* 6(2), 155-168.

Rahman, M. and L. M. Pearson (2000). "Shapiro-Francia W' Statistic Using Exclusive Simulation", *Journal of the Korean Data & Information Science Society,* 11(2), 139-155.

Rahman, M. (1999). "Estimating the Box-Cox Transformation via Shapiro-Wilk W Statistic." *Communications in Statistics - Simulation and Computation,* 28(1), 223-241.

Shapiro, S. S. and M. B. Wilk (1965). "An Analysis of Variance Test for Normality." *Biometrika,* 52, 3 and 4, 591-611.

Shapiro, S. S., M. B. Wilk, and H. J. Chen (1968). "A Comparative Study of Various Tests of Normality." *Journal of the American Statistical Association,* 63, 1343-1372.

Taylor, J. M. G. (1985). "Power Transformations to Symmetry." *Annals of Mathematical Statistics,* 33, 1-67.

Weisberg, S. (1985). *Applied Linear Regression, 2nd ed.* John Wiley & Sons, New York.