Academic Journal of Nawroz University (AJNU), Vol.12, No.4, 2023

https://doi.org/10.25007/ajnu.v12n4a1708

Develop a Nonlinear Regression Model Using Box-Cox Transformation with Application

Rozheen Taher Awdi [1], Haithem Taha Mohammed Ali [2,3]

[1] Department of Statistics, University of Duhok, Kurdistan Region, Iraq
[2] Department of Economic Sciences, University of Zakho, Kurdistan Region, Iraq
[3] Department of Economics, Nawroz University, Kurdistan Region, Iraq

ABSTRACT

This article introduces an algorithm designed for utilizing power transformations in the estimation of nonlinear regression models. The algorithm outlines a series of steps for selecting the most suitable power parameter estimate through a combination of the conventional Maximum Likelihood Estimation technique and specific criteria for enhancing statistical modeling effectiveness. Supplementary decision guidelines involve the utilization of the determination coefficient and the p-value from the errors normality test. The algorithm's application was demonstrated using actual data. The article's conclusion highlighted the ability to identify a range of feasible solutions for selecting the optimal power parameter. However, it was acknowledging the challenge of identifying a single optimal value that satisfies the requirements of all estimation and decision methodologies.

KEY WORDS: Multiple linear regression, Box-Cox transformation, Power parameter.

1. INTRODUCTION

The fundamental prerequisites for statistical inference in the estimation and testing of the Multiple Linear Regression (MLR) model encompass the normality and constancy of variance in the estimated model errors [1]. Consequently, techniques for data transformation, especially those within the domain of power transformations, have been harnessed to significantly amplify the effectiveness of statistical modeling and achieve an improved overall fit. The approach of employing the Box-Cox Transformation (BCT) was dedicated to meeting the modeling prerequisites of the MLR Model by applying parametric power transformations [2]. In 1955, Tukey introduced a set of power transformations in which the transformed values displayed a monotonic relationship with the observations across a permissible range [3]. In 1974, Box and Hill unveiled their approach involving power transformations and specialized weights, designed to attain homogenous variance properties—a technique they termed the "weighted power transformation" [4]. In 1983, Atkinson proposed a transformation method that addresses anomalies in data post-transformation [5]. Sakia's 1992 study revisited BCT, aiming to simplify the model and align the theoretical assumptions with more satisfactory analyses [6]. Cook and Weisberg's 1994 method aimed to find a linear and monotonic transformation of the response variable based on the BCT model [7]. Yeo and Johnson, in 2000, introduced a distribution family that retained the beneficial properties of BCT while removing constraints on its usage, including the handling of positive and negative values [8]. Hossain, M. Z., in 2011, provided an analytical review of BCT's significant role across various statistical domains, including estimation, testing, inference, and model selection [9]. Al-Yousef and Abduahad, in 2014, developed a technique using power transformations to describe Bayesian conditional expectation probability distributions within a nonlinear regression framework [10]. Fischer, in 2016, proposed logarithmic transformations and tree growth models for left-skewed variable transformation, catering to regression analysis assumptions [11]. Soleymani's 2018 modification of BCT aimed to enhance estimation accuracy, particularly in econometrics and time series applications [12]. In 2021, Atkinson, Riani, & Corbellini extended BCT to non-negative responses in linear regression models, encompassing both-sided transformations and the Yeo-Johnson transformation for observations with positive or negative values [13].

This article's objective is to present an algorithm that leverages BCT to construct a nonlinear multiple regression model when the prerequisites for MLR analysis are not satisfied. The algorithm takes into account the multitude of criteria for estimating the optimal power parameter value. The subsequent sections are structured as follows: The second section delves into theoretical facets of BCT. The third section outlines the algorithm proposed for developing a nonlinear regression model through BCT. The fourth section focuses on practical implementation. The fifth and final section encompasses the concluding remarks.

2. BOX-COX TRANSFORMATION

In 1964, Box and Cox introduced a transformation model of significant importance within the realm of statistics [2]. They put forth a pair of techniques for estimating the power parameter in the Tukey transformation model. The primary approach involves Maximum Likelihood Estimation (MLE), while the secondary method adopts a Bayesian perspective. The overarching objective of the BCT is to mitigate anomalies in data, address nonlinearity, rectify errors' non-normality, and manage heteroscedasticity [2]. BCT is given by,

$$\psi(y) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases} \tag{1}$$

Where Y is the response variable and λ is the power parameter. However, the most common transformations of BCT are described in Table 1.

Table 1: The most common transformations of BCT

λ	Transformation
-2	$1/y^2$
-1	$1/y$
-0.5	$1/\sqrt{y}$
0	ln (y)
0.5	$\sqrt{y}$
1	Y
2	y^2

In the case of one-dimensional data, the methodology assumes for any random variable Y, if $Z = \psi(y)$ represent the transformed variable of Y such that $Z \sim N(\mu, \sigma^2)$, then the probability density function (PDF) of the random variable Y is given by $f_Y(y; \lambda, \mu, \sigma^2) = f_Z(\psi^{-1}(y); \lambda, \mu, \sigma^2) \,.\, J(\lambda, y)$. Thus, the criteria for choosing the optimal estimator of λ is to maximize the log likelihood of the PDF of the original observations except for a constant,

$$L_{\max}(\lambda, y) = -\,(n/2)\log \sigma^2\,(\lambda) + \log\ J(\lambda, y) \tag{2}$$

Where $\hat{\sigma}^2\ (\boldsymbol{\lambda})$ is the variance estimator of Z. In the case of the MLR model defined as,

$$\mathbf{Z} = \mathbf{X\beta} + \mathbf{e} \tag{3}$$

Where, $\mathbf{Z} = \boldsymbol{\psi}(\mathbf{y})$ represents the (nx1) column vector of the transformed values of the response variable vector, **X** is the (nxp) known information matrix, **β** is the (px1) unknown parameters vector and **e** represent the (nx1) column vector of random errors and distributed according to the normal distribution with means vector equal to (nx1) zero vector and identity variances matrix equal to $\sigma^2\mathbf{I_n}$. The assumption of errors normality feature leads to the fulfillment of a normality too in transformed response data vector **Z** according to the following joint PDF,

$$f_{\mathbf{Z}}(\mathbf{z}; \lambda, \mathbf{X\beta},\ \sigma^2) = (2\pi\pi\sigma^2)^{-n/2}.\exp\left\{\frac{-(\mathbf{Z}-\mathbf{X\beta})^T(\mathbf{Z}-\mathbf{X\beta})}{2\sigma^2}\right\}\ ,\mathbf{Z} \in R \tag{4}$$

Using the change of variables technique, the following equations system represents the joint PDF of original response data vector,

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi\sigma^2)^{-n/2}.\exp\left\{\frac{-(\boldsymbol{\psi}^{-1}(\mathbf{y})-\mathbf{X\beta})^T(\boldsymbol{\psi}^{-1}(\mathbf{y})-\mathbf{X\beta})}{2\sigma^2}\right\}.\left|\frac{d\psi(\mathbf{y})}{d\mathbf{y}}\right| \tag{5}$$

Similarly, in the scenario involving a single variable, the criterion for selecting the optimal estimator for λ aims to maximize the log-likelihood of the joint probability density function of the original observations, excluding a constant term,

$$L_{\max}(\lambda, y) = -\,(n/2)\log \hat{\sigma}^2(\lambda) + \log J(\lambda, y) \qquad (6)$$

When $\mathbf{Y}$ is replaced by $\psi(\mathbf{y})$ for some λ, the back transformation of BCT is of the form [12],

$$\mathbf{Y} = \begin{cases} (\lambda\,\psi(\mathbf{y}) + 1)^{1/\lambda} & \text{if} \quad \lambda \neq 0 \\ \exp(\psi(\mathbf{y})) & \text{if} \quad \lambda = 0 \end{cases} \qquad (7)$$

Hence, upon estimating the MLR using the transformed data, we can deduce an estimation of the original data's nonlinear model through the subsequent equation for estimated back transformation,

$$\hat{\mathbf{y}} = \begin{cases} (\lambda\,\mathbf{X}\boldsymbol{\beta}\hat{} + 1)^{1/\lambda} & \text{if} \quad \lambda \neq 0 \\ \exp(\mathbf{X}\boldsymbol{\beta}\hat{}) & \text{if} \quad \lambda = 0 \end{cases} \qquad (8)$$

3. ALGORITHM

This article presents a novel application algorithm that leverages the BCT model alongside parametric estimation to construct a nonlinear multiple regression model. The process of determining the optimal power parameter λ within this algorithm rests upon three distinct criteria:

First, the algorithm considers $L_{\max}(\lambda, \mathbf{y})$, which involves the MLE of the PDF for the original random variable Y as defined in Eq. 6 [2].

The second criterion revolves around maximizing explanatory model efficiency, specifically by aiming for the highest Coefficient of Determination R^2 value within the estimated MLR applied to the transformed random variable Z, according to Eq. 3. For a more comprehensive understanding of the array of criteria for optimal power parameter selection, refer to [14] and [15].

The third criterion involves selecting the highest p-value from the Shapiro-Wilk test for the errors' normality resulting from the estimated nonlinear regression model of the original data vector, as specified by Eq. 8 [16].

Consequently, the proposed application algorithm for implementing the BCT model and parametric estimation to construct a nonlinear multiple regression model is structured as follows:

Step 1: Estimate the MLR model of $Y/X_1, X_2, \dots, X_k$.

Step 2: Fix $\lambda \in \Lambda$, where $\Lambda = \{-2, -1.9, \dots, 0, \dots, 1.9, 2\}$

Step 3: Transform Y to $Z = \psi(y)$ using BCT according to Eq. 1.

Step 4: Estimate MLR model of the transformed data vector $Z/X_1, X_2, \dots, X_k$ according to Eq. 3 and the coefficient of determination R^2 for all $\lambda \in \Lambda$.

Step 5: Estimate the values of MLE according to Eq. 6 for all $\lambda \in \Lambda$.

Step 6: Estimate the nonlinear multiple regression model for the original data vector according to back transform Eq. 8.

Step 7: Calculate the P-value of Shapiro-Wilk test of the errors vector normality of the estimated nonlinear model of the original data vector resulting from step 6.

Step 8: Repeat all the steps from 3 to 7 for all values of λ in Λ.

4. APPLICATION

The BCT was applied to the cellphone dataset, and an R program was used to analyze the data. The data can be found at (https://www.kaggle.com/datasets/mohannapd/mobile-price-prediction). The Cellphone dataset has a size of 161 observations which contains one dependent variable Y representing the price and the following twelve independent variables; sale, weight, resolution, pixels per inch, central processing unit core (CPU core), central processing unit frequency (CPU freq.), internal memory, the random access memory (RAM), rear camera, front camera, charge and battery, and the thickness.

Through the utilization of the presented algorithm, we achieved the generation of a convex curve representing the MLE function employing equation (6). The peak value of this curve corresponds to the most advantageous power parameter value. As illustrated in Figure 1, the calculated power parameter value stood at 0.8. In Table 2 and Figures 2 and 3, can observe the estimations and patterns of the Coefficient of Determination R^2 in accordance with Step 4, along with the P-value stemming from the Shapiro-Wilk test evaluating the normality of the errors vector, as outlined in Step 7.

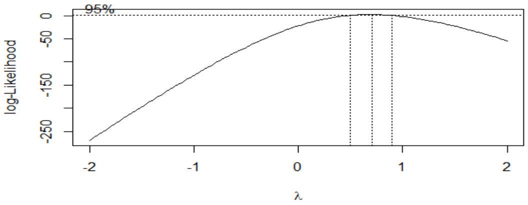

Figure 1: Log-likelihood plot according to the step 5

Table 2: The estimates of R^2 according to Step 1, p-value of Shapiro-Wilk test of the errors vector normality according to Step 7

λ	-2	-1.9	-1.8	-1.7	-1.6	-1.5	-1.4	-1.3	-1.2	-1.1
R^2	0.84	0.85	0.86	0.86	0.87	0.88	0.89	0.89	0.90	0.90
P-value	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00
λ	-1	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
R^2	0.91	0.92	0.92	0.93	0.93	0.94	0.94	0.94	0.95	0.95
P-value	0.00	0.14	0.00	0.21	0.21	0.01	0.04	0.23	0.01	0.01
λ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
R^2	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96
P-value	0.01	0.10	0.01	0.23	0.09	0.17	0.24	0.23	0.20	0.16
λ	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
R^2	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.93	0.92
P-value	0.29	0.27	0.21	0.14	0.08	0.16	0.01	0.00	0.23	0.09

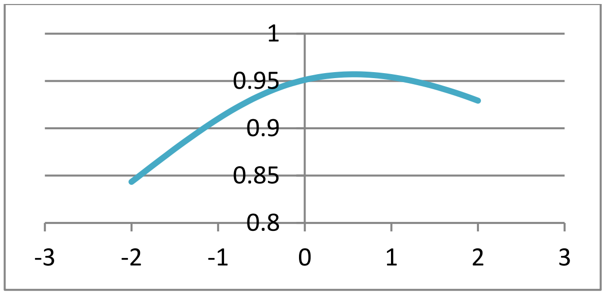

Figure 2: The plot of the Coefficient of Determination R^2 according to step 4

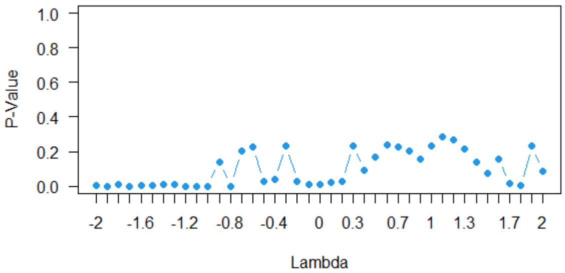

Figure 3: The plot of the p-value of Shapiro-Wilk test of the errors vector normality according to Step 7

Within our algorithm, we have derived three distinct criteria for the selection of the most suitable transformation parameter value. It became evident to the authors that achieving a single optimal value that aligns with the requirements of all three criteria – the maximum MLE function value, R^2, and the p-value from the Shapiro-Wilk test assessing the normality of the errors vector – is improbable. Frequently, a reassessment of outcomes becomes necessary, considering the significance and precedence of particular criteria and how other criteria might contribute to reinforcing these priorities. The estimations of the optimal power parameter, as per the three criteria, were tabulated in Table 3.

Table 3: The estimates of the optimal power parameter according to the three criteria

Optimum **λ**	Criteria values		
	L_{max}	R^2	p-value
0.8	0	0.96	0.2048
0.6	-5	0.96	0.2365
1.1	-5	0.95	0.2885
1.0	-7	0.95	0.2336

From table 3, a feasible solution of the optimal value of the power parameter can be deduced in the ranges of $L_{max}(\lambda, \mathbf{y}) \epsilon (-7,0)$, $R^2 \epsilon (0.95, 0.96)$ and $p - \text{value}\ \epsilon\ (0.2048, 0.2885)$. If we consider that L_{max} the MLE value of PDF of the original random variable Y according to Eq. 6, represents the basis for estimating the optimum power parameter, then the optimal value is 0.8 because it fulfills at least one of the other two criteria .

$$\hat{y} = (0.8\,\mathbf{X}\widehat{\boldsymbol{\beta}} + 1)^{1/0.8} \qquad (9)$$

Where **X** is the (nxp) known information matrix and β is the following (px1) parameters estimates vector,

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 465.56 \\ -0.001 \\ -0.17 \\ -8.48 \\ 0.21 \\ 11.28 \\ 31.58 \\ 1.22 \\ 19.59 \\ 1.26 \\ 1.94 \\ 0.03 \\ -16.37 \end{bmatrix}$$

5. CONCLUSION

Applying power transformation to reshape the response variable in regression contexts offers a pathway to constructing a nonlinear model when the prerequisites of linear regression analysis fall short. Various methodologies exist for selecting the most appropriate power parameter, organized into two categories: the first involves well-established estimation techniques like the MLE approach, while the second employs efficiency criteria within regression modeling as decision-making guidelines for power parameter determination.

This article embarked on the task of defining a workable range where multiple estimation techniques and decision rules converge, enabling the identification of the optimal parameter that best satisfies a multitude of efficiency-enhancing requirements in regression modeling. If we consider the MLE value $L_{\max}(\lambda, \mathbf{y})$ related to the original random variable Y, as depicted in Eq. 6, it forms the basis for deriving the optimal power factor, then, the other criteria can then function as assisting factors in the selection of the optimal power parameter.

REFERENCES

[1] Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (Eds.). (1998). Applied regression analysis: a research tool. New York, NY: Springer New York.

[2] Box, G. E., & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological), 26(2), 211-243.

[3] Tukey, J. W. (1957). On the comparative anatomy of transformations. The Annals of Mathematical Statistics, 602-632.

[4] Box, G. E., & Hill, W. J. (1974). Correcting inhomogeneity of variance with power transformation weighting. Technometrics, 16(3), 385-389.

[5] Atkinson, A. C. (1983). Diagnostic regression analysis and shifted power transformations. Technometrics, 25(1), 23-33.

[6] Sakia, R. M. (1992). The Box-Cox transformation technique: a review. Journal of the Royal Statistical Society: Series D (The Statistician), 41(2), 169-178.

[7] Cook, R. D., & Weisberg, S. (1994). Transforming a response variable for linearity. Biometrika, 81(4), 731-737.

[8] Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87(4), 954-959.

[9] Hossain, M. Z. (2011). The use of Box-Cox transformation technique in economic and statistical analyses. Journal of Emerging Trends in Economics and Management Sciences, 2(1), 32-39.

[10] Alyousif, H. T., & Abduahad, F. N. (2014). Develop a Nonlinear Model for the Conditional Expectation of the Bayesian Probability Distribution (Gamma–Gamma). Al-Nahrain Journal of Science, 17(2), 205-212.

[11] Fischer, C. (2016). Comparing the logarithmic transformation and the box-cox transformation for individual tree basal area increment models. Forest Science, 62(3), 297-306.

[12] Soleymani, S., (2018). Exact Box-Cox Analysis. A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree in Doctor of Philosophy in IJSAS, The University of Western Ontario.

[13] Atkinson, A. C., Riani, M., & Corbellini, A. (2021). The box–cox transformation: Review and extensions.

[14] Draper N. R. & and Smith H., (1998). “Applied Regression Analysis,” *John Wiley Sons.*, vol. 326, doi: 10.1080/00401706.1967.10490452.

[15] Othman, S. A., & Ali, H. T. M. (2021). Improvement of the nonparametric estimation of functional stationary time series using Yeo-Johnson transformation with application to temperature curves. Advances in Mathematical Physics, 2021, 1-6.

[16] Hou, Q., Mahnken, J. D., Gajewski, B. J., & Dunton, N. (2011). The Box-Cox power transformation on nursing sensitive indicators: Does it matter if structural effects are omitted during the estimation of the transformation parameter? BMC Medical Reserch Methodology, 11,1-12.