ISSN 1533-9211

DOI: 10.5281/zenodo.10622840

A COMPARISON OF DIFFERENT CRITERIA TO CONSTRUCT REGRESSION MODEL EMPLOYING THE BOX-COX AND COLE GREEN TRANSFORMATION

AZAD ADIL SHAREEF

Department of Statistics, College of Administration and Economics, University of Duhok, Kurdistan Region, Iraq.
Email: azada@uod.ac

Abstract

This article introduces an algorithm that utilizes power transformation to estimate a nonlinear regression model for Cole Green and Box-Cox transformation. The algorithm outlines steps for selecting an optimal powers parameter estimate, employing the Akaike Information Criterion and Bayesian Information Criterion, statistical modeling efficiency criteria are incorporated to complement the traditional method. Additionally, Decision rules include the adjusted coefficient of determination Maximum Likelihood Estimator and the F-statistics test. The proposed algorithm is applied to real data, and the conclusion emphasizes the feasibility of obtaining various options exist for selecting the optimal power parameter. However, attaining a singular optimal value that meets both estimation and decision criteria methods is deemed impractical.

Keywords: Cole Green Transformation, Box-Cox Transformation, Adjusted R-Square, and Akaike Information Criterion and Bayesian Information Criterion.

1. INTRODUCTION

The essential requirements in statistical inference for testing and estimating the multiple linear regression (MLR) model, it is essential to verify the constancy of variance and normality in the estimated model errors. [1]. Hence, the transformed data aimed at achieving linearity, particularly those within the family of power transformations has been employed to significantly improve the effectiveness involving statistical modeling, with the overarching objective of achieving a better fit. The Box-Cox transformation (BCT) approach was specifically chosen to meet the modeling conditions in MLR by employing a parametric power transformation [2]. In 1992, Sakia, R. M. conducted a study on the revision of Box-Cox transformation (BCT), with a focus on streamlining the model and identifying a scale that aligns more closely with the theoretical assumptions made in the analysis, thereby enhancing the overall satisfaction with the model [3]. In 1994, Cook and Weisberg introduced a technique designed to identify a linear and monotonic transformation of the dependent variable, adhering to the BCT model [4]. In 2000, Yeo, I. K., and Johnson, R. A. introduced a novel family of distributions that can be applied without constraints, possessing several favorable properties akin to the Box-Cox transformation (BCT). Their extension of BCT forms a single-parameter family, permitting its use in scenarios involving both positive and negative variable values [5]. In 2011, Hossain conducted an analytical review highlighting the substantial role of the BCT, methodology applied across diverse statistical domains, encompassing estimation, and testing [6]. In 2021, Atkinson, Riani, and Corbellini focused on the BCT applied to non-negative responses within linear regression models. The discussed extensions involve transforming both

The Seybold Report

ISSN 1533-9211

DOI: 10.5281/zenodo.10622840

sides of the model, as well as the utilization of the Yeo-Johnson transformation for observations that can exhibit either positive or negative values. [7]. In 2022, authors Al-Safar and Mohammed Ali employed power transformations to enhance nonlinear models within the framework of Response Surfaces Methodology [8].

The aim of this article is to propose an algorithm to compare different criteria to develop a nonlinear multiple regression model for employing the Box-Cox and Cole Green transformation

To estimate the optimal value of the power parameter. The rest of the article is organized as follows: The second section includes some theoretical aspects about the criteria and transformation models. The third section includes the proposed algorithm to develop a nonlinear regression model using BCT and CGT. The fourth section includes practical aspect of the article. While the fifth section includes the conclusions.

2. CRITERIA AND TRANSFORMATION MODELS

In statistics, AIC, Represented by Akaike Information Criterion, this metric gauges the comparative quality of statistical models. applied to a given dataset. The AIC is often used for model selection, where you have several candidate models and you want to determine which one is the most appropriate for describing the underlying structure in your data [9]. One of the widely adopted information criteria is AIC. The concept of AIC, introduced by Akaike in 1998, involves selecting the model that minimizes the negative likelihood penalized by the number of parameters, as defined in the Eq.(1) [10]. The AIC is calculated based on the likelihood function of the model and penalizes models for their complexity. The concept involves striking a balance between the goodness of fit and the simplicity of the model. A lower AIC is indicative of a better-performing model.

The formula for AIC is given by [11]:

$$AIC = -2 \times log\, p(L) + 2p \qquad (1)$$

Where L refers to the likelihood under the fitted model and p Represents the count of parameters within the model. The model with the lowest AIC is generally preferred, as it suggests the best trade-off between goodness of fit and simplicity.

Bayesian Information Criterion (BIC) is a criterion used for model selection in statistics. Both AIC and BIC are measures of the goodness of fit of a statistical model, but they incorporate a penalty for the number of parameters in the model to avoid overfitting [12].

The BIC is calculated using the following formula:

$$BIC = -2 \times log\, p(L) + p \;\; log\,(n) \qquad (2)$$

The penalty term ($p \;\; log\,(n)$) in BIC is larger than in AIC and is proportional to the logarithm of the sample size. The purpose of the penalty term is to discourage overly complex models, especially when dealing with small sample sizes. The BIC tends to favor simpler models compared to AIC.

When comparing models using BIC, the model with the lowest BIC is considered the best-fitting model. Like AIC, BIC helps strike a balance between model fit and model complexity, but it tends to be more conservative in selecting simpler models, particularly when the sample size is small Assessing the impact of data structure on the accuracy of the estimators R-square and adjusted R-square in multiple linear regression using Monte Carlo simulation involves generating synthetic data sets with different structures and assessing how well the regression models perform in terms of R-square and adjusted R-square [13]. The adjusted R^2_{adj} is a modified version of the regular coefficient of determination (R^2) in the context of linear regression models [14]. It penalizes the inclusion of unnecessary predictors in the model, addressing the issue of overfitting.

The formula for R^2_{adj} is given by:

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1} \qquad (3)$$

Where n is the number of observations and k is the number of predictors in the model.

The F-test is typically used for assessing the equality of variances among groups, commonly in the context of Analysis of Variance (ANOVA). On the other hand, a power transformation (possibly referring to a transformation) is often used to address issues like non-normality or heteroscedasticity in the data [15]. The F-test is commonly used to compare the variances of two or more groups. In ANOVA, for example, the F-test is used to determine if there are statistically significant differences in means among groups. The assumption of equal variances is important in ANOVA, and the F-test is employed to assess this assumption.

In 1964, Box and Cox introduced a pivotal transformation model in statistics, providing two approaches for estimating to obtain the power parameter. The method incorporates MLE, followed by employing a Bayesian approach. The goal of the BCT is to rectify anomalies in data, alleviate nonlinearity, address non-normality of errors, and counteract heteroscedasticity. The BCT formula is expressed as follows:

$$\psi(y) = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda} & if \quad \lambda \neq 0 \\ ln(y) & if \quad \lambda = 0 \end{cases} \qquad (4)$$

In 1992, Cole and Green transformation (CGT), represented Y as dependent variable, is presumed to be positive. Assuming Y possesses a median denoted by $\boldsymbol{\mu}$, and when raised to the power of y^{λ} , or if $\lambda = 0$, the $ln(y)$ follows a normal distribution. In such cases, it is suitable to examine the transformed variable [16].

$$\psi(y) = \begin{cases} \dfrac{\left(\frac{y}{\mu}\right)^{\lambda}-1}{\lambda} & if\ \lambda \neq 0 \\ ln\left(\frac{y}{\mu}\right) & if\ \lambda = 0 \end{cases} \qquad (5)$$

derived from the transformation family introduced by Box and Cox. This mapping transforms the medianμ of Y to $\psi(y) = 0$, and is continuous at $\lambda = 0$.

Derived from the transformation family introduced by Box and Cox, this conversion aligns the median μ of Y to $\psi(y)$= 0 and maintains continuity at $\lambda = 0$.

In the MLR model, Y represents the response variable, the methodology presupposes specific conditions for any random variable Y, if $W = \psi(y)$ illustrate the transformed variable of Y such that $W \sim N(\mu, \sigma^2)$, then the probability density function (PDF) of the random variable Y is given by $f_Y(y; \lambda, \mu, \sigma^2) = f_W(\psi^{-1}(y); \lambda, \mu, \sigma^2) .J(\lambda, y)$. Therefore, the criteria for selecting the optimal estimator of λ involve maximizing the log-likelihood of the probability density function (PDF) of the original observations, excluding a constant term.

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

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

$$\boldsymbol{W} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (7)$$

Where, $\boldsymbol{W} = \boldsymbol{\psi}(\boldsymbol{y})$, the expression “represents $(n \times 1)$ vector column comprising the altered values of the response variable vector” indicates that the given variable is a column vector of size $(n \times 1)$ where n represents the number of elements. This column vector contains the transformed values of the response variable vector. $\boldsymbol{X}$ is the $(n \times p)$ known information matrix. $\boldsymbol{\beta}$ is the $(p \times 1)$ unknown parameters vector, and $\boldsymbol{\varepsilon}$ represent the $(n \times 1)$ The "column vector of random errors" signifies a vector containing random error values. Additionally, it is mentioned that these errors are distributed in accordance with the normal distribution, and their means vector is equal to a certain value. The complete statement would depend on the specific context and details of the means vector $(n \times 1)$, zero vector and identity variances matrix equal to $\sigma^2 \boldsymbol{I}_n$. The assumption of normality in the errors features results in the transformed response data vector $\boldsymbol{W}$ also exhibiting normality. This is in accordance with the following joint Probability Density Function (PDF),

$$f_W(\boldsymbol{w}; \lambda, \boldsymbol{X\beta},\ \sigma^2) = (2\pi\sigma^2)^{-n/2}.\ exp\left\{\frac{-(\boldsymbol{W} - \boldsymbol{X\beta})^T(\boldsymbol{W} - \boldsymbol{X\beta})}{2\sigma^2}\right\} \qquad , \boldsymbol{W}\ \epsilon\ R \qquad (8)$$

By utilizing the change of variables method, the subsequent system of equations illustrates the joint. PDF of the original response data vector:

$$f_Y(\boldsymbol{y}) = \ (2\pi\sigma^2)^{-n/2}.\ exp\left\{\frac{-(\boldsymbol{\psi}^{-1}(\boldsymbol{y}) - \boldsymbol{X\beta})^T(\boldsymbol{\psi}^{-1}(\boldsymbol{y}) - \boldsymbol{X\beta})}{2\sigma^2}\right\} . \left|\frac{d\psi(\boldsymbol{y})}{d\boldsymbol{y}}\right| \qquad (9)$$

In the context for a single variable, the method of choosing the optimal estimator (λ) involves maximizing the logarithm of the joint probability density function (PDF) of the original observations, excluding a constant., When Y is replaced by their BCT and CGT to $\psi(\mathbf{y})$ for some λ, making the back transform of BCT and CGT, respectively to get,

The Seybold Report

ISSN 1533-9211

DOI: 10.5281/zenodo.10622840

$$\boldsymbol{Y} = \begin{cases} (\lambda\,\psi(\boldsymbol{y}) + 1)^{1/\lambda} & if \qquad \lambda \neq 0 \\ exp\big(\psi(\boldsymbol{y})\big) & if \qquad \lambda = 0 \end{cases} \tag{10}$$

and

$$\boldsymbol{Y} = \begin{cases} ((\lambda\,\psi(\boldsymbol{y}) + 1)\mu^{\lambda})^{1/\lambda} & if \qquad \lambda \neq 0 \\ exp\big(\psi(\boldsymbol{y})\big)\mu & if \qquad \lambda = 0 \end{cases} \tag{11}$$

Therefore, upon estimating the Multiple Linear Regression (MLR) of the transformed data, we can derive an estimation for a nonlinear representation of the original data model using the following back-transformed equations.

$$\hat{\boldsymbol{y}} = \begin{cases} (\lambda\,\boldsymbol{X}\hat{\boldsymbol{\beta}} + 1)^{1/\lambda} & if \qquad \lambda \neq 0 \\ exp\big(\boldsymbol{X}\hat{\boldsymbol{\beta}}\big) & if \qquad \lambda = 0 \end{cases} \tag{12}$$

and

$$\hat{\boldsymbol{y}} = \begin{cases} ((\lambda\,\boldsymbol{X}\hat{\boldsymbol{\beta}} + 1)\mu^{\lambda})^{1/\lambda} & if \qquad \lambda \neq 0 \\ exp\big(\boldsymbol{X}\hat{\boldsymbol{\beta}}\big)\mu & if \qquad \lambda = 0 \end{cases} \tag{13}$$

3. ALGORITHM

In this article, the author has presented an algorithm that utilizes the BCT and CGT model, along with parametric estimation, to build a multiple regression model.. The selection of the optimal power parameter λ in this algorithm relies on five distinct criteria; AIC, BIC, Adjusted R^2, F-statistics and MLE. Hence, the outlined application algorithm involves the utilization of the BCT and CGT model and parametric estimation for the development of multiple regression model, and it proceeds as follows:

Step 1: Estimate the (MLR) model for the given data $Y/X_1, X_2, \dots, X_6$.

Step 2: Fix λ ∈ Λ, where $\Lambda = \{-b, -b + 0.1, \dots, b - 0.1, b\}$

Step 3: Estimate the value of AIC according to Eq. (1) for all λ ∈ Λ.

Step 4: Estimate the value of BIC according to Eq. (2) for all λ ∈ Λ.

Step 5: Transform Y to W = ψ(y) using CGT and BCT according to Eq. (4) and Eq. (5).

Step 6: Estimate MLR model of the transformed data vector $\mathrm{W}/\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_6$ according to Eq. (7) and the Adjusted R^2 according to Eq. (3) for all λ ∈ Λ.

Step 7: Estimate the values of F- statistics for MLR model of the transformed data vector $\mathrm{W}/\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_6$ for all λ ∈ Λ.

Step 8: Compute the MLE values using Eq. (6) for all λ in the set Λ.

Step 9: Repeat the procedures from step 2 to step 8 for each λ within the set Λ.

ISSN 1533-9211

DOI: 10.5281/zenodo.10622840

4. APPLICATION

The Cellphone dataset underwent BCT and CGT model for analysis using R program. The dataset, accessible at https://www.kaggle.com/datasets/mohannapd/mobile-price-prediction, comprises 161 observations. It encompasses a dependent variable, denoted as Y denotes the price, accompanied by six independent variables: sale, weight, resolution, pixels per inch, central processing unit core (CPU core), and central processing unit frequency (CPU freq.). Our algorithm has identified five distinct criteria for choosing the optimal value of the transformation parameter for CGT and BCT (see figure 1 and figure 2).

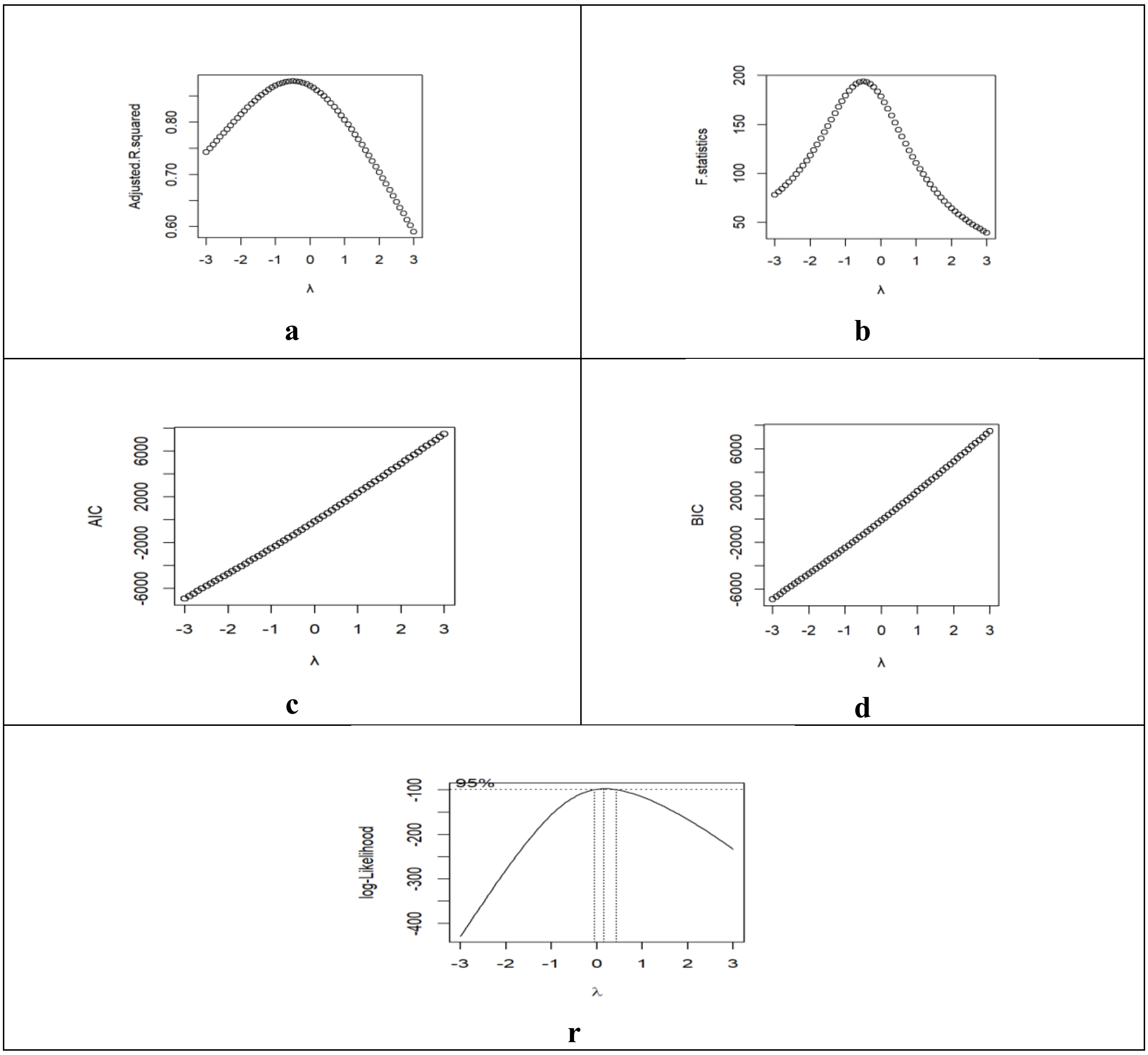

Figure 1: For all λ ∈ Λ (a) The adjusted R^2 (b) The values of F-statistics (c)The values of AIC (d) The values of BIC (e) Log-likelihood curve for BCT

DOI: 10.5281/zenodo.10622840

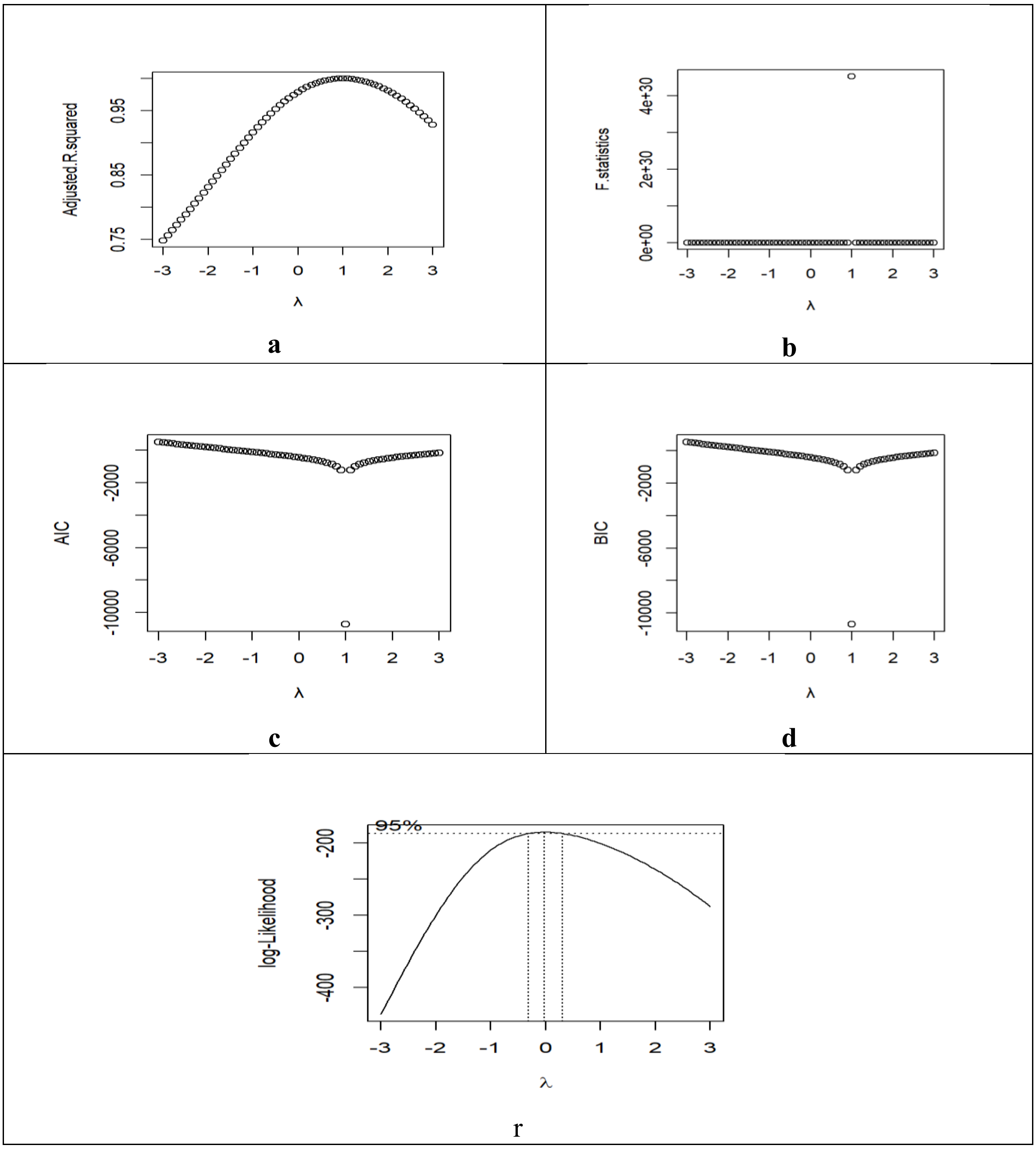

Figure 2: For all $\lambda \in \Lambda$ (a) The adjusted R^2 (b) The values of F-statistics (c) The values of AIC (d) The values of BIC (e) Log-likelihood curve for CGT

It can be seen that from proposed algorithm, the author obtained the convex curve of the MLE according to Eq. (6) for CGT and BCT, the optimal value of the power parameter corresponds to the peak of this curve that can be seen that in figure 1 and figure 2, the estimated value of the power parameter is 0.2 and 0, respectively.

ISSN 1533-9211

DOI: 10.5281/zenodo.10622840

Table 1 shows the output for all values $\lambda \in \Lambda$ for five different criteria which are Adjusted R^2, F-statistics, AIC, BIC, and MLE for both models. In numerous instances, it is necessary to assess the results based on the significance and priority of certain criteria, considering the additional support that other criteria can offer to these priorities. Table 2 presents the estimations of the optimal power parameter according to the five criteria.

Table 1: The estimatation of all criteria according to algorithm steps for both model BCT and CGT

λ	BCT					CGT				
	Adjusted R^2	F-statistics	AIC	BIC	MLE	Adjusted R^2	F-statistics	AIC	BIC	MLE
-3	0.74	78.02	-6858.7	-6834.0	-428.7	0.75	69.07	509.57	537.30	-436.41
(-2.9, -2)	(0.75, 0.81)	(81, 118.4)	(-6643, -4700)	(-6619, -4675)	(-410, -283)	(0.75, 0.83)	(72, 113)	(199, 478)	(227, 505)	(-418, -303)
(-1.9, -1)	(0.82, 0.87)	(123.9, 179.4)	(-4482, -2492)	(-4458, -2467)	(-266, -155)	(0.84, 0.91)	(121, 251)	(-74, 168)	(-77, 196)	(-288, -216)
(-0.9,-0.1)	(0.87, 0.88)	(183.8, 194)	(-2265, -390)	(-2240, -366)	(-143, -99)	(0.93, 0.97)	(279, 875)	(-403, -105)	(-375, -108)	(-200, -184)
0	0.87	178.6	-148.58	-123.93	-98.75	0.98	1067	-441.96	-414.23	-184.74
(0.1, 0.9)	(0.81, 0.87)	(110.6, 172.5)	(95, 2089)	(120, 2113)	(-115, - 97.17)	(0.98, 0.99)	(1329, 1.14E+05)	(-1222, -441)	(-1194, -455)	(-197, -184)
1	0.80	110.6	2342.45	2367.10	-115.34	0.99	4.53E+30	-10711.9	-10684.2	-200.42
(1.1, 2)	(0.70, 0.79)	(64.37, 104.7)	(2596, 4901)	(2620, 4925)	(-164, -117)	(0.98, 0.99)	(1186, 1.16E+05)	(-1223, -454)	(-1195, -426)	(-235, -203)
(2.1, 3)	(0.59, 0.69)	(41.35, 61.14)	(5159, 7492)	(5183, 7517)	(-233, -171)	(0.92, 0.97)	(296, 981)	(-384, -157)	(-390, -129)	(-287, -240)

Table 2 displays the optimal values of λ for each indicator when in its optimal state.

The Adjusted R^2 for the original dataset is 0.80 which increases to 0.88 when power parameter becomes -0.5 if we compare it with the original dataset for BCT but there is no improvement in the CGT. The F-statistics for the original dataset is 110.6 which increases to 194 when power parameter becomes -0.5 if we compare it with the original dataset for BCT but there is no improvement in the CGT. The minimum value for both criteria AIC and BIC is -6858.68 and -6834.02 for BCT when power parameter becomes -3 if we compare it with the original dataset. However, the minimum value for both criteria AIC and BIC is -10711.9 and -10684.2 for CGT when power parameter becomes 1 that means this is an original dataset. If $\mathrm{L}_{\mathrm{max}}$ denotes the MLE value of PDF for the original random variable Y, serving as the foundation for estimating the optimal power parameter, the values of MLE for CGT and BCT are -184.742 and -97.17, respectively. the optimal value is identified as 0 and 0.2 for CGT and BCT respectively. As a result, the researcher infers that the MLE function displays a convex curve, as illustrated in Table 2. Therefore, the researcher reaches the conclusion that the MLE function exhibits a convex curve. Ultimately, in this particular application, the researcher determined that selecting the optimal value for the power parameter was achievable using MLE for CGT and BCT of the estimated linear regression of the transformed response vector for dataset.

Table 2: The estimations of the optimal power parameter in accordance with the five criteria for both methods

Criteria according to BCT	Values	Optimal Power Parameter (λ)	Criteria according to CGT	Values	Optimal Power Parameter (λ)
Adjusted R^2	0.88	-0.5	Adjusted R^2	0.99	1
F-statistics	194	-0.5	F-statistics	4.53E+30	1
AIC	-6858.68	-3	AIC	-10711.9	1
BIC	-6834.02	-3	BIC	-10684.2	1
MLE	-97.17	0.2	MLE	-184.742	0

5. CONCLUSION

The criteria used for model selection clearly pinpoint the appropriate asymmetric model among various competing alternatives. Essentially, the findings underscore the significance of design characteristics when conducting studies on asymmetric cellphone transmission. In unstable conditions, such as situations with small sample sizes, five different criteria was used for both model CGT and BCT. AIC demonstrates superior performance compared to BIC. The comparison presented adds to our knowledge and comprehension of the relative efficacy of AIC and BIC within an asymmetric cellphone transmission modeling framework, an area that has been relatively underexplored. The confirmation of the validity of AIC and BIC in selecting the appropriate model for cellphone transmission in the current studies implies that other estimators based on AIC and BIC may also prove promising as model selection criteria. Various techniques exist for determining the optimal power parameter. Hence, the researcher concludes that the MLE function demonstrates a convex curve. In this specific application, the researcher determined that the selection of the optimal power parameter value was attainable through

The Seybold REPORT

ISSN 1533-9211

DOI: 10.5281/zenodo.10622840

MLE for both CGT and BCT in the estimated linear regression of the transformed response vector for the dataset and the researcher found that the BCT is better than CGT to obtain the optimal power parameter. This article aimed to explore a viable solution space for several methods of estimation and rules for decision-making in order to identify the optimal parameter that fulfills the maximum number of efficiency improvement criteria for regression modeling.

References

1) D. Wang and M. Murphy, "Estimating Optimal Transformations for Multiple Regression Using the ACE Algorithm," *J. Data Sci.*, vol. 2, no. 4, pp. 329–346, 2021, doi: 10.6339/jds.2004.02(4).156.

2) J. Abrevaya, "Computing marginal effects in the box–cox model," *Econom. Rev.*, vol. 21, no. 3, pp. 383–393, 2002, doi: 10.1081/ETC-120015789.

3) Sakia Remi M., "The Box-Cox transformation technique: a review," *J. R. Stat. Soc. Ser. D (The Stat.*, vol. 41, no. 2, pp. 169–178, 1992.

4) R. D. Cook and S. Weisberg, "Transforming a response variable for linearity," *Biometrika*, vol. 81, no. 4, pp. 731–737, 1994.

5) Yeo I.K. and Johnson R.A., "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 156, no. I, pp. 87–90, 2000.

6) Hossain M. Z., "The use of Box-Cox transformation technique in economic and statistical analyses," *J. Emerg. Trends Econ. Manag. Sci.*, vol. 2, no. 1, pp. 32–39, 2011.

7) Atkinson A.C. and R.M. and C.A., *The Box–Cox Transformation: Review and Extensions*, vol. 36, no. 2. 2021, pp. 239–255. doi: 10.1214/20-STS778.

8) A. Al-Saffar and H. T. M. Ali, "Using Power Transformations in Response Surface Methodology," in *2022 International Conference on Computer Science and Software Engineering (CSASE)*, 2022, pp. 374–379.

9) H. Bozdogan, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

10) H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected papers of hirotugu akaike*, Springer, 1998, pp. 199–213.

11) H. de-G. Acquah, "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship," 2010.

12) K. A. Bollen, J. J. Harden, S. Ray, and J. Zavisca, "BIC and alternative Bayesian information criteria in the selection of structural equation models," *Struct. Equ. Model. a Multidiscip. J.*, vol. 21, no. 1, pp. 1–19, 2014.

13) A. Y. J. Akossou and R. Palm, "Impact of data structure on the estimators R-square and adjusted R-square in linear regression," *Int. J. Math. Comput*, vol. 20, no. 3, pp. 84–93, 2013.

14) H. Piepho, "An adjusted coefficient of determination (R2) for generalized linear mixed models in one go," *Biometrical J.*, p. 2200290, 2023.

15) F. Nwobi and F. Akanno, "Power comparison of ANOVA and Kruskal–Wallis tests when error assumptions are violated," *Adv. Methodol. Stat.*, vol. 18, no. 2, pp. 53–71, 2021.

16) T. J. Cole and P. J. Green, "Smoothing reference centile curves: the LMS method and penalized likelihood," *Stat. Med.*, vol. 11, no. 10, pp. 1305–1319, 1992.