Iraqi Journal of Statistical Sciences

https://stats.uomosul.edu.iq

Analyzing the Impact of Various Criteria on Blood Pressure Through General Linear Model by Box-Cox Transformations

Ziba Mohammed Mustafa [1*] Azad Adil Shareef [2]

[1,2] Department of Statistics and Informatics, College of Administration and Economics, University of Duhok, Kurdistan Region, Iraq.

Article information

Article history:
Received: June 15,2025
Revised: July 31,2025
Accepted: September 15,2025
Available online: June 1,2026

Keywords:
Box–Cox transformation, statistical criteria, general linear model and Shapiro–Wilk test

Correspondence:
Ziba Mohammed Mustafa[1*]
ziba.mustafa@uod.ac
Azad Adil Shareef [2]
Azad.shareef@uod.ac

Abstract

Blood pressure is a key indicator of cardiovascular health, reflecting the force exerted by blood against arterial walls. Data transformation tools are essential in statistical analysis for improving assumptions necessary for linear models, such as normality, linearity, and homoscedasticity, especially when these assumptions are violated. These techniques are especially helpful for correcting distortions in data structure and enhancing the validity of General Linear Models (GLMs). In this study, we applied the Box-Cox transformation, a method from the family of power transformations, to improve a linear regression model. Our objective was to identify the most appropriate power transformation to enhance model performance and interpretability, using statistical criteria on blood test datasets collected from Azadi Hospital in Duhok. These evaluation criteria are essential for the accurate interpretation of data, as they assess the model's quality and reliability. The criteria used included: adjusted R-squared, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), F-statistic, Maximum Likelihood Estimation (MLE), Root Mean Square Error (RMSE), and the Shapiro–Wilk test. These measures also guided the selection of the most appropriate GLM. We can conclude that different λ values optimize different model performance aspects: $\boldsymbol{\lambda = 0.7}$ balances good model fit ($Adj.R^2$, F-statistic). $\boldsymbol{\lambda = -3}$ gives the most accurate predictions (lowest error). $\boldsymbol{\lambda = 1.3\ or\ 1.5}$ improves likelihood and residual normality. A computational algorithm was proposed to estimate the optimal power parameter, and the results of the criteria were discussed and compared. Based on the adjusted R-squared criterion, an optimal λ value was identified, indicating a strong model fit.

DOI: 10.33899/iqjoss.v23i1.62125 ,

1. Introduction

The General Linear Model (GLM) has evolved over the past two centuries, beginning with the least squares method introduced by Legendre (1805) and later extended by Gauss (1809), who incorporated the assumption of normally distributed errors. Nelder and Wedderburn (1972) expanded the framework to handle non-normal response variables through the development of the general linear model, significantly broadening its applicability. The GLM represents a broad class of models used to describe the relationship between a continuous response variable and one or more explanatory (predictor) variables, encompassing techniques such as simple linear regression, multiple regression, and ANOVA. The general linear model relies on key assumptions: the response variable is normally distributed, the relationship between the response and predictors is linear, and the error variance (σ^2) is constant across all levels of the predictors (Dobson, 2002).

Data transformation is important in constructing linear models, as it is used to fix major model assumptions when they are violated. Ruppert (2001) states that the careful application of transformations can improve both model fit and interpretability, and that in many cases, transforming the predictors is less impactful than transforming the response. Early transformations such as logarithmic transformation, square root, and reciprocal transformations were introduced to stabilize variance and reduce skewness. A major advancement in data transformation came with the Box–Cox transformation (BCT). Box and Cox (1964) made a significant contribution to regression modeling by introducing a structured method for transforming the dependent variable to better meet the assumptions of the general linear model. In regression analysis, the validity of inference relies heavily on assumptions. One of the key insights from their work is that transforming the dependent variable alone is often sufficient to correct issues like non-linearity or heteroscedasticity, as transformations of the independent variables usually have a limited effect on improving model assumptions.

A limitation of the Box-Cox family of transformations is that they can only be applied to positive data. Yeo and Johnson (2000) extend the Box-Cox family to allow for both zero and negative values of the response variable. This method improves the normality and symmetry of data distributions and is especially useful in regression modeling when the data contain non-positive values. The authors demonstrate that the transformation maintains the flexibility and interpretability of Box-Cox while offering broader applicability and improved robustness in practical data analysis. Yang (2006) proposes a new class of power transformations, the Dual Power Transformation (DPT), as a robust alternative to the Box-Cox method, specifically designed to address its limitations, most notably the truncation issue when handling zero or negative values. He also demonstrates that the proposed transformations can lead to improved model fit and more accurate statistical inference. Since then, several alternatives have been proposed, and modern methods now incorporate statistical tests to evaluate transformation effectiveness. Data transformation remains a critical step in ensuring the validity of statistical inference in linear modeling.

In this study, multiple linear regression was employed to analyze blood test data and assess the impact of several independent variables on a continuous outcome variable. Hou et al. (2011) evaluated the effectiveness of the Box-Cox transformation in analyzing nursing-sensitive indicators and concluded that the Box-Cox transformation improves model fit and inference, particularly in complex healthcare data with hierarchical or repeated measures, and remains effective even when some structural factors are unknown. Raymaekers and Rousseeuw (2020) propose a novel framework for transforming data; their method focuses on robustly normalizing only the central portion of the data, reducing sensitivity to outliers and extreme values. This approach improves the reliability of statistical analyses, particularly in regression and machine learning contexts where normality assumptions are important for model validity and inference. Hawkins (n.d.) critically evaluates the common practice of assuming normality after transformation without verification and recommends conducting formal goodness-of-fit tests on the transformed data. This step is essential to ensure the reliability of statistical inferences, particularly in regression analysis, where the assumption of normally distributed error terms is fundamental.

Wichitaksorn et al. (2014) develop a generalized class of skew distributions to support robust quantile regression modeling. Their approach addresses common challenges in regression analysis, such as skewness, heavy tails, and outliers, which often violate the assumptions of classical linear regression. By incorporating flexible error distributions into the quantile regression framework, the authors enhance the model's ability to capture asymmetries in the conditional distribution of the response variable, which is especially useful in financial and time series applications. Pek et al. (2017) caution against simplistic inverse transformations and recommend only using transformations when they enhance interpretability or meet specific inferential goals, especially in small-sample contexts where the central limit theorem may not suffice. Transformations not only alter the scale of the transformed variables but may also change the fundamental relationships among variables while simultaneously affecting the distribution of model errors. In medical and health research, data often violate the assumptions of normality and homoscedasticity required for many statistical analyses. Power transformations are essential tools for: Normalizing skewed clinical data (e.g., blood pressure, viral load), stabilizing variance across groups, improving linear relationships in regression models, enabling the application of parametric tests.

The focus of this study is to show the importance of BCT on blood test data to improve the relationship between variables and enhance model performance. Since the response variable in this case consists solely of positive values, the application of the Box-Cox transformation is both appropriate and valid. We employed various statistical criteria and compared them to obtain the best value of transformation parameter (λ) in the R programming language across multiple measures and found that the BCT impacts improving model performance.

This study aims to analyze the impact of various demographic, behavioral, and physiological criteria on blood pressure using a General Linear Model enhanced by Box-Cox transformations. By applying this approach, the study seeks to ensure the statistical validity of inferences and improve the interpretability and predictive performance of the model.

This methodology not only accounts for the skewed nature of blood pressure data but also provides a robust foundation for identifying key factors that contribute to its variation in the population.

2. Materials and methods

2.1. Diabetes dataset

Blood pressure is one of the most critical indicators of heart health, with elevated levels serving as a primary risk factor for heart disease, and kidney failure. As such, understanding the factors that influence blood pressure—such as age, body mass index (BMI), gender, lifestyle, and socioeconomic variables—is a major focus in medical and public health research. However, analyzing blood pressure data poses several statistical challenges, including non-normal distributions, heteroscedasticity, and the presence of outliers. These issues can explore the assumptions of traditional linear models, potentially leading to biased or inefficient estimates. The data for this study were collected from the laboratory of Azadi Hospital in Duhok City, which contains blood test results from 221 individuals. The attributes of the dataset are as follows: Hemoglobin (HGB) is a protein molecule in red blood cells responsible for carrying oxygen; platelets (PLT) are small cell fragments in the blood that help stop bleeding; age refers to the age of the individual; white blood cells (WBC) are immune system components responsible for defending the body against infections; red blood cells (RBCs) carry oxygen to the body's tissues from the lungs; and mean corpuscular hemoglobin concentration (MCHC) represents the typical hemoglobin content of red blood cells relative to their volume.

2.2. Methods

In general, transformation methods are used to improve the assumptions required for constructing an appropriate statistical model. In this study, we developed a multiple linear regression model to examine the relationship between various blood indicators, with hemoglobin (HGB) levels as the dependent variable. Since the data consisted of strictly positive blood test results, the Box-Cox transformation was applied to the response variable to identify the optimal power transformation. The model parameters were estimated using Maximum Likelihood Estimation (MLE). To evaluate model performance across different transformations, several statistical criteria were computed, including Adjusted R-squared, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), F-statistic, MLE values, and the Shapiro–Wilk test for normality. The final selection of the optimal value of λ was based primarily on the Adjusted R-squared criterion.

2.3. Multiple Linear Regression

The first algorithm we implemented was a multiple linear regression model, a statistical method used to examine the relationship between multiple independent variables and a single dependent variabl. Marill (2004) states that multiple linear regression allows the investigator to account for all of these potentially important factors in one model. The advantages of this approach are that it may lead to a more accurate and precise understanding of the association of each factor with the outcome. It also plays a critical role in clinical research by allowing investigators to control for variables that could bias the results, thus strengthening the validity of conclusions drawn from observational data. Lin et al. (2023) note that multiple linear regression analysis can produce prediction results that are more accurate by creating a linear relationship model between variables. To investigate the relationship between various predictors and the factors influencing hemoglobin (HGB) levels, which served as the dependent variable, the model included five independent variables: PLT, age, WBC, RBC, and MCHC. As a result, according to the coefficient of determination (Adjusted R-squared), approximately 76.52% of the variability in HGB levels is explained by these independent variables. This indicates that the selected variables are significant contributors to the model and suggests a strong model fit, meaning the model explains a substantial portion of the variance in HGB levels, which is an indication of a robust and reliable regression model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad i = 1, 2, 3, \dots, n \tag{1}$$

Where:

y_i is the dependent variable.

$x_1, x_2, \dots, x_i$ are the independent variables.

β_0 is the intercept term,

$\beta_1, \beta_2, \dots, \beta_k$ are the regression coefficients corresponding to each independent variable,

ε_i is the error term (representing the residuals).

We estimate the parameters of the model by using the maximum likelihood method. Accurate parameter estimation enables reliable predictions and meaningful interpretations of the effects of explanatory variables.

2.4. Maximum Likelihood Estimation (MLE)

In the second algorithm, model parameters were estimated using Maximum Likelihood Estimation (MLE). In a multiple linear regression model, each parameter estimate (β) represents the expected change in the dependent variable for a one-unit increase in the corresponding independent variable, assuming all other variables are held constant. Ramachandran and Tsokos (2009) discussed this parameter estimation technique, proposed by statistician and geneticist Sir Ronald A. Fisher in 1922. It finds the values of the parameters that make the observed data most probable and performs well with large sample sizes. Mohammed and Mahdi (2025) stated that the MLE method plays a crucial role in diagnosing the minimum random error thresholds for probability distributions, thereby supporting the identification of optimal parameter values and the assessment of the distribution's behavior. In the context of linear regression, MLE provides estimates for the regression coefficients β by assuming that the errors are normally distributed, with constant variance. In this analysis, all parameter estimates were statistically significant, with $p-values$ less than 0.05. This indicates strong evidence that each independent variable has a meaningful association with the dependent variable (HGB levels). In other words, a significant β suggests that the variable contributes to explaining the variability in HGB levels and should be retained in the model. Conversely, if a coefficient is not significant, it implies that the associated variable has little or no linear impact on the dependent variable in the presence of the other predictors. The Box–Cox transformation was applied to satisfy model assumptions and enhance the relationship between the variables and the model.

After applying the Box-Cox transformation, we calculated the MLE criteria for all λ values, with the results shown in appendix (A) Table (1), and illustrated in appendix (B) Figure 1(e), titled "Lambda with MLE" this plot highlights how different values of λ impact the MLE, making it a valuable tool for selecting the optimal transformation parameter. Observing the pattern of black dots in Figure 1(e), we note that the MLE reaches its peak at approximately $\lambda = 1.3$, indicating the most likely transformation under the assumed distribution. As λ moves away from this value in either direction, the likelihood decreases, suggesting a poorer fit. This analysis helps identify the transformation parameter that maximizes the likelihood of the transformed data being normally distributed, which is crucial for improving model assumptions and overall fit.

$$f(y_i \mid x_i, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}\right) \quad (2)$$

$$E(y) = \mu \qquad Y \sim N(\mu, \sigma^2)$$

$$f(y_i \mid x_i, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-1}{2\sigma^2}(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}))^2\right) \quad (3)$$

The likelihood function is the pdf viewed as a function of the parameters β, σ^2:

$$L(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma^2 \mid Y) = \prod_{i=1}^{n} f(y_i \mid \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma^2) \quad (4)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})\right)^2\right] \quad (5)$$

Take the Log of the Likelihood (Log-Likelihood):

$$= \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})\right)^2 \quad (6)$$

when dealing with multivariate linear regression and estimating parameters by minimizing the sum of squared errors, it's often clearer and more elegant to rewrite the expression using techniques of Numerical analysis.

$$Y_{(n\times1)} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X_{(n\times k)} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \qquad \beta_{(n\times1)} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$$

Then the model becomes:

$$\boldsymbol{Y} = \boldsymbol{X\beta} \tag{7}$$

The objective function (sum of squared errors) becomes:

$$LL(\beta, \sigma^2 \mid Y) = \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}[(Y - X\beta)^T(Y - X\beta)] \tag{8}$$

$$= \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}[Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta] \tag{9}$$

The derivative of equation (5) with respect to β is, and then equal to zero

$$\frac{dLL(\beta, \sigma^2 \mid Y)}{d_\beta} = -\frac{1}{2\sigma^2}(2X^TX\beta - 2X^TY) \tag{10}$$

$$X^TY - X^TX\beta = 0 \tag{11}$$

$$\hat{\beta} = (X^TX)^{-1}\, X^TY$$

2.5. The Box-Cox Transformation (BCT)

The Box-Cox Transformation is a parametric power transformation developed by George Box and David Cox (1964). It is widely used to make the response variable more normally distributed, stabilize variance (achieve homoscedasticity), improve linearity between predictors and the response, and enhance the validity of statistical inference in linear regression or ANOVA models.

The third algorithm we applied was the Box-Cox Transformation (BCT) to the dependent variable HGB levels to improve the normality, linearity, and homoscedasticity of linear regression model. The Box-Cox Transformation, introduced by George Box and Sir David Roxbee Cox in 1964, is a family of power transformations that is widely used for transforming dependent variables. It is designed to stabilize variance and make the data conform more closely to a normal distribution. It is only applicable to strictly positive response values and does not accommodate zero or negative values. The Box-Cox transformation is a crucial method in linear regression when a regression model's assumptions are seriously violated. The transformation is based on a power parameter λ and influences the kind of transformation the response variable is subjected to. We considered λ to be a member of a set Λ and changed its value in an interval from -3 to 3 with a step of 0.1, where $\Lambda = \{-3, -2.9, -2.8, \ldots, 2.9, 3\}$. We extended the range of λ slightly beyond the standard or commonly used range to allow more flexibility in exploring the best transformation. For each value of λ, the transformed model was fitted using Maximum Likelihood Estimation (MLE), and several metrics were computed. The goal was to identify the λ value that maximized model performance and best satisfied the assumptions of linear regression after the transformation.

The transformation is defined as a function of a parameter λ, and the transformed response variable $Y^{(\lambda)}$is given by:

$$y(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & ; \lambda \neq 0 \\ \log(y) & ; \lambda = 0 \end{cases} \tag{12}$$

Atkinson *et al* (2021) $\lambda = 1$ denotes no transformation, when $\lambda = 1$, the Box–Cox formula simplifies to $y^{(1)} = y - 1$. Although this appears as a constant shift, in regression modeling, the intercept term absorbs such shifts in the response. Therefore, $\lambda = 1$ is treated as the "no transformation" case, since the scale, distribution, and the correlation between variables remain unchanged, $\lambda = 1/2$ to the square root transformation, $\lambda = 0$ to the logarithmic transformation, and $\lambda = -1$ to the reciprocal transformation, in order to prevent a discontinuity at zero.

2.6. Criteria

In this section we applied the following criteria in this study.

2.7.1 Adjusted R-squared

The fourth algorithm involved calculating the Adjusted R-squared for all values of λ. Adjusted R-squared is a statistical measure used in regression analysis to determine the proportion of variance in the dependent variable explained by the independent variables while adjusting for the number of predictors in the model. According to Wohlwend (2023), when you have multiple regression models with varying numbers of predictors, you should use Adjusted R-squared, which is a modified version of R-squared that has been adjusted for the number of predictors in the model. Pirenne and Claeskens (2024) state that in linear models, model selection is frequently based on this criterion, which calculates the percentage of response variance that can be accounted for by the covariates while controlling for model complexity. Miles (2014) calculated it using:

$$Adjusted\ R^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-p-1}\right) \quad (13)$$

Where:
R^2: coefficient of determination
n: number of observations
p: number of predictors (independent variables)

Adjusted R-squared can’t be negative. R-squared, which measures the proportion of explained variance, is always between 0 and 1 because it is based on the ratio of sums of squares, which are always non-negative. A negative estimate of variance would be statistically meaningless.

In this algorithm, the optimal power parameter λ yields the highest Adjusted R-squared value. Higher values of Adjusted R-squared indicate a better model fit. The results are presented in Table 1, and Figure 1(a) illustrates how the criteria values change in response to variations in λ. The horizontal axis of the plot represents the range of λ values, varying from -3 to 3. As observed, the Adjusted R-squared increases steadily as λ moves from highly negative values toward zero. The curve reaches its peak around $\lambda = 0.7$, where the Adjusted R-squared attains its maximum value of 0.7666. This point represents the optimal balance between model complexity and goodness of fit. Beyond this value, as λ continues to increase, the Adjusted R-squared begins to decline, suggesting that the model becomes overly simplified and starts to underfit the data.

2.7.2 Root Mean Square Error (RMSE)

Chai and Draxler (2014) state that the square root of the Mean Square Error (MSE) gives a measure of the mean magnitude of the prediction errors, expressed in the same units as the dependent variable. Wohlwend (2023) notes that it is useful to utilize when you want to provide a more interpretable explanation of a model's performance. In general, a lower RMSE value denotes smaller variances between predicted and observed values, which is indicative of superior model performance. It tells us how far, on average, the predicted values are from the actual values with larger errors penalized more heavily due to squaring.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (14)$$

Where:
n is how many observations there are in the dataset.
y_i is the true value of the observation.
$\hat{y}_i$ is the expected value of the i^{th} observation.

The fifth algorithm we estimated the (RMSE) using the Box-Cox Transformation (BCT) for all values of λ and selected the optimal power parameter corresponding to the lowest RMSE. The RMSE values are presented in Table 1, and the trend is illustrated in Figure 1(d). This plot shows how RMSE changes in response to variations in λ. The black dots in the plot display an upward trajectory, indicating a direct relationship between λ and prediction error: as λ increases, the RMSE also increases. This trend suggests that higher values of λ degrade the model’s predictive accuracy, resulting in less precise predictions. Based on this criterion, the optimal value of the power parameter λ is -3, where the prediction errors are minimized, indicating the best model performance under this transformation.

2.7.3 Mean Absolute Error (MAE)

Chai and Draxler (2014) and Wohlwend (2023) state that MAE, like MSE and RMSE, measures the difference between actual and expected values, giving equal weight to all errors. This makes it a useful measure for determining the average error magnitude without overly penalizing large errors.

The Formula

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{15}$$

Where:
n is how many observations there are in the dataset.
y_i is the true value of the observation.
$\hat{y}_i$ is the expected value of the i^{th} observation.

The sixth algorithm we estimated the Mean Absolute Error (MAE) using the Box-Cox Transformation (BCT) for all values of λ. The MAE values are presented in Table 1, and the trend is illustrated in Figure 1(c). This plot shows how the MAE changes in response to variations in λ. The black dots in the plot display an upward trajectory, indicating a direct relationship between λ and prediction error: as λ increases, the MAE also increases. This trend suggests that higher values of λ degrade the model's predictive accuracy, resulting in less precise predictions. Based on this criterion, the optimal value of the power parameter is $\lambda = -3$, where the prediction errors are minimized, indicating the best model performance under this transformation.

2.7.4 Mean Absolute Percentage Error (MAPE)

MAPE measures regression model accuracy, especially in time series, by averaging the absolute percentage differences between the projected and actual values. It helps compare performance across datasets or models and is simple to understand. According to de Myttenaere et al. (2016), the use of MAPE is pertinent to finance, where relative measures are frequently used to quantify gains and losses. Lower MAPE values indicate higher predictive accuracy.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100\% \tag{16}$$

Where:
n is the number of observations in the dataset.
y_i is the actual value of the observation.
$\hat{y}_i$ is the predicted value of the i^{th} observation.

The seventh algorithm we calculated MAPE, as well as other criteria, for all values of λ, with the optimal power parameter corresponding to the lowest MAPE. The MAPE values are presented in Table 1, and the trend is illustrated in Figure 1(b). This plot shows how MAPE changes in response to variations in λ. The black dots in the plot display an upward trajectory, indicating a direct relationship between λ and prediction error: as λ increases, the MAPE also increases. This trend suggests that higher values of λ degrade the model's predictive accuracy, resulting in less precise predictions. Based on this criterion, the optimal value of the power parameter is $\lambda = -3$, where the prediction errors are minimized, indicating the best model performance under this transformation.

2.7.5 F-Statistic

Gupta and Kapoor (2000), F is the proportion of two independent chi-square variates divided by their respective degrees of freedom, and the result follows Snedecor's F-distribution with v degrees of freedom. Seber and Lee (2003) state that the F-statistic is used for testing hypotheses, making it possible to determine whether the model is appropriate. El-Horbaty (2024) discusses its application in some important unbalanced linear mixed models.

$$F = \frac{SSR/v_1}{SSE/v_2} = \frac{MSR}{MSE} \tag{17}$$

Where:
SSR is the sum of squares regression, with $v1 = p$ degree of freedom, where p is the number of coefficients in the model.
SSE is sum of squares error, with $V2 = N - p - 1$ degree of freedom, where N is the total sample size.

A large F-statistic indicates that the model explains a substantial amount of the variance and is likely a good fit. To determine statistical significance, the F-statistic is compared to a critical value from the F-distribution or evaluated using a p-value.

The eighth algorithm we estimated the F-statistic and the optimal power parameter corresponding to the highest F-statistic. The results are presented in Table 1 and illustrated in Figure 1(f), which shows how the F-statistic changes as λ varies. Figure 1(f) that the F-statistic increases as λ moves from -3 to 0, reaching its peak around $\lambda = 0$, and then gradually decreases as λ continues to increase toward 3. This trend indicates that the model achieves its strongest statistical significance at $\lambda = 0.7$, consistent with the results reported in Table 2. As λ deviates further from this value, the F-statistic declines, suggesting a reduction in the model's explanatory power and overall effectiveness. Therefore, $\lambda = 0.7$ appears to provide the optimal balance between model fit and statistical significance based on the F-statistic criterion.

2.7.6 Shapiro–Wilk Test

The ninth algorithm we assessed the normality of the residuals after transformation from the linear regression model by applying the Shapiro–Wilk test across all values of λ. One test of normality is the Shapiro-Wilk test, which was published by Samuel Sanford Shapiro and Martin Wilk in 1965. Khatun (2021), focusing on the importance of normality as a fundamental assumption in many statistical analyses, including linear regression, ANOVA, and hypothesis testing, also emphasizes that violations of normality can lead to incorrect inferences, biased estimates, and reduced statistical power. A p-value below the alpha level suggests non-normality, while a higher p-value means normality cannot be rejected. A statistical model's goodness of fit indicates how well the model represents a collection of observations and is often used to compare the performance or explanatory power of different models.

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (18)$$

Where:
W: The Shapiro-Wilk statistic
n: sample size
$y_{(i)}$: the i-th order statistic (i.e., the i_th smallest value in the sample)
$\bar{y}$: the sample mean
a_i: constants derived from the means, variances, and covariance of the ordinary normal distribution's order statistics
The numerator represents the squared correlation between the order statistics and the expected values under normality.
The denominator is the sample variance.

The test results are summarized in Table 1 and visualized in Figure 1(g), titled "Lambda with Shapiro. Test," which illustrates the relationship between λ and the Shapiro–Wilk test statistic. This statistic, which ranges from approximately 0.75 to 0.975, is used to evaluate the normality of a dataset, with values closer to 1 indicating better adherence to a normal distribution. As λ increases from -3 to approximately 1, the Shapiro–Wilk statistic also increases, suggesting an improvement in the normality of the residuals. The test statistic reaches its peak around λ = 1, after which it slightly declines, indicating a modest reduction in normality. This plot provides valuable insight into how different λ values affect the distributional properties of the transformed data. Since the objective is to identify the λ value that best normalizes the data, $\lambda \approx 1.5$ appears to be the most appropriate choice based on the Shapiro–Wilk test results.

3. Result and Discussion

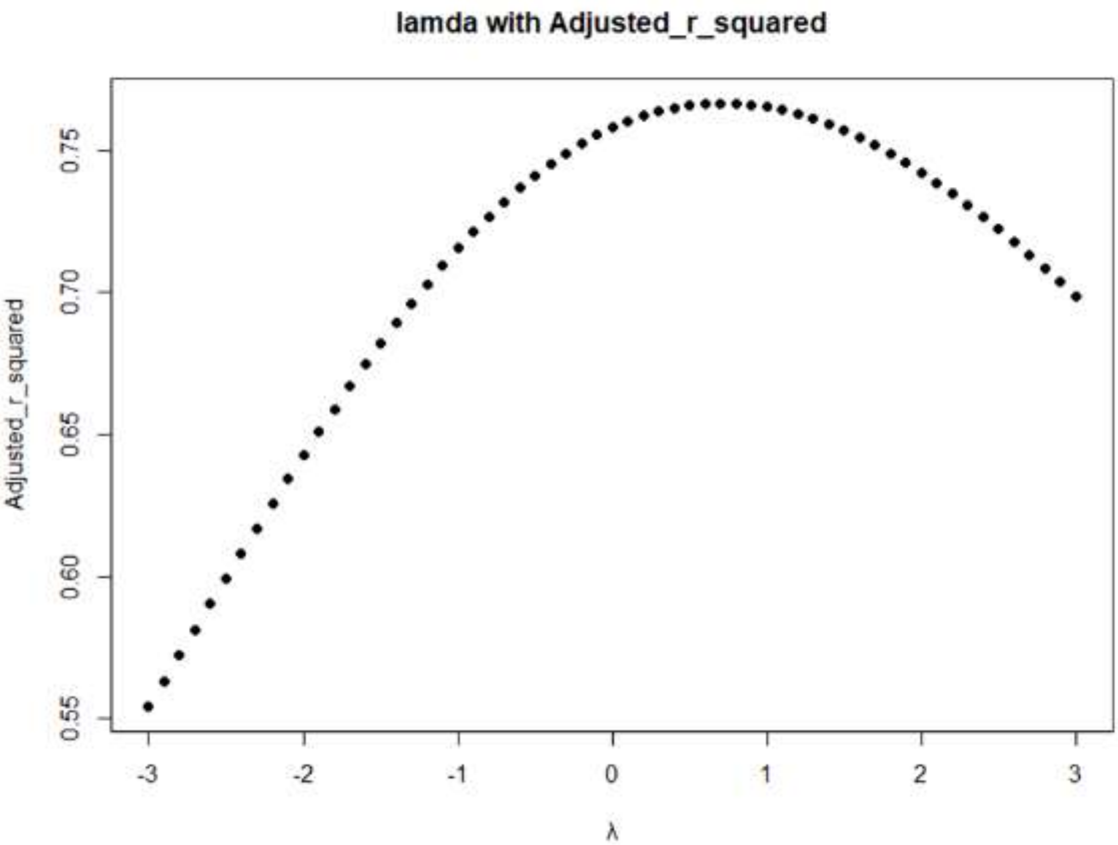
Each row represents a range of power transformation parameters (-3, -2.9, ..., 3). For each range, Table 1 provides several performance and fit criteria. Normality (SW) Increases as the power value approaches 1, peaking around 1.5 to 2 (0.975,0.977). This suggests that the residuals become more normal around these values. Prediction Error (MAPE, MAE, RMSE) Errors (MAE, RMSE) are lowest in the range (-3, -2) and steadily increase as λ increases beyond 0. In contrast, MAPE increases gradually across the same range. MLE is estimated as the most negative values (lowest MLE) occur for more negative λ, indicating better likelihood fits earlier in the range. However, this should be balanced with other criteria. F-statistic Peaks at $\lambda = (0.6, 0.9)$, meaning the model has maximum explanatory power around this range. Table 2 compares different power parameters (λ) used in the Box-Cox transformation to determine which value best optimizes the performance of a General Linear Model (GLM) applied to blood pressure data, each row shows how a specific λ value performs according to a different model evaluation criterion. Adjusted R^2 Indicates the proportion of variance in blood pressure explained by the model after adjusting for the number of predictors. When $\lambda = 0.7,$ it yields the best model fit in terms of explained variance. MAPE, MAE, and RMSE are error metrics indicating how close model predictions are to actual values. Lower values indicate better accuracy; when $\lambda = -3,$ it minimizes all three-error metrics, suggesting it provides the most accurate predictions. The maximum likelihood value from fitting the model fewer negative values are better; for example, $\lambda = 1.3$ gives the best likelihood, indicating it may be the best overall in terms of model fit from a probabilistic standpoint. The F-statistic, measures the overall significance of the model. A higher F-statistic ($at\ \lambda = 0.7$) suggests the predictors are highly significant under this transformation. The Shapiro–Wilk (SW) Test tests for normality of residuals; values closer to 1 indicate more normal residuals. When $\lambda = 1.5,$ it gives the most normally distributed residuals, supporting the assumptions of GLM. We can conclude that different λ values optimize different model performance aspects: $\lambda = 0.7$ balances good model fit ($Adj.R^2$, F-statistic), $\lambda = -3$ gives the most accurate predictions (lowest error), $\lambda = 1.3\ or\ 1.5$ improves likelihood and residual normality. In this study, the optimal value of the BCT parameter λ was determined based on the best value of the adjusted R^2 criteria. The primary objective of the transformation is to improve model assumptions and enhance predictive accuracy at the optimal λ.

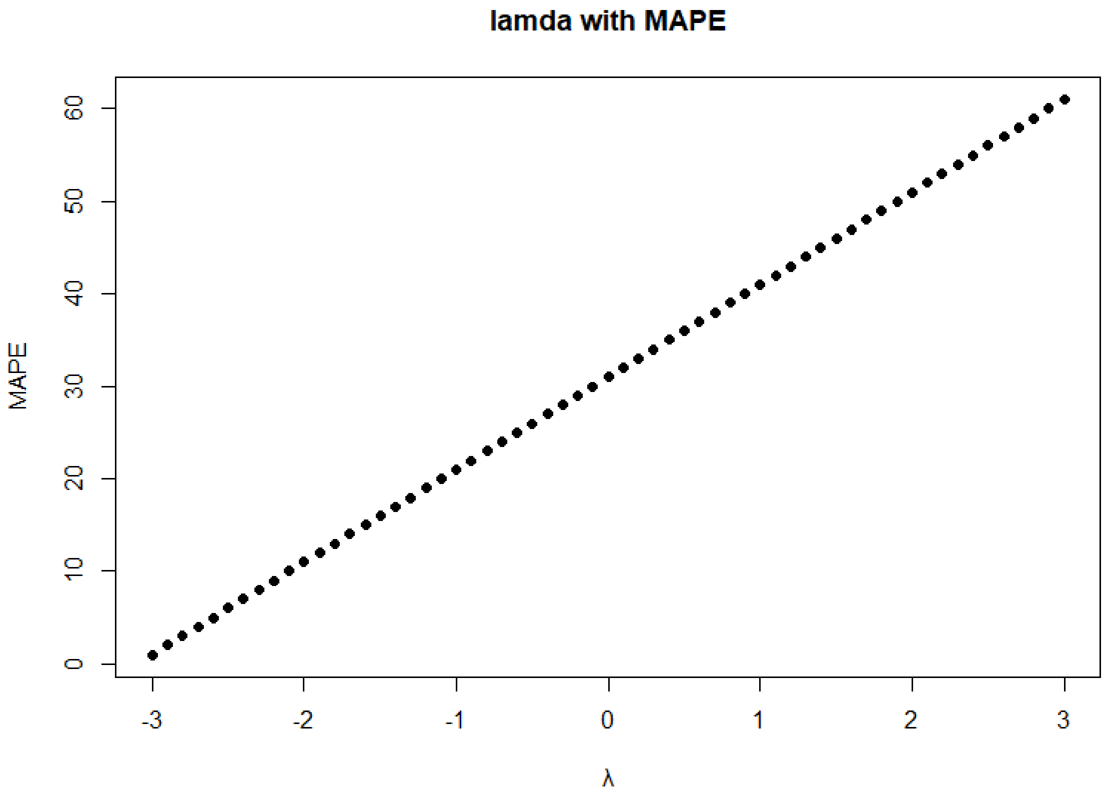
Table1: Calculated Criteria for Selecting Optimal Power Transformation

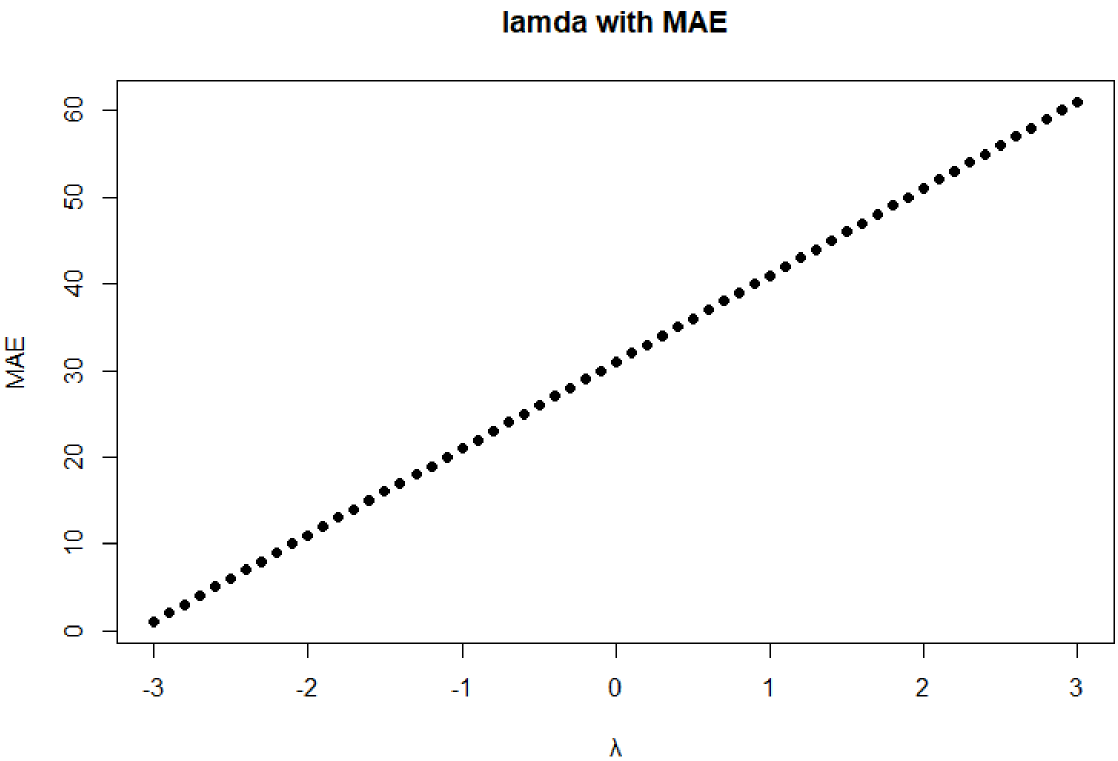
λ rang	***Adj.R²***	***MAPE***	***MAE***	***RMSE***	***MLE***	***F-statistic***	***SW***
(-3, -2.5)	(0.554,0.599	(0.020, 0.050)	(6.735E-05 0.00019)	(0.00011, 0.00032)	(-278.066, -235.519)	(55.71, 66.83)	(0.754, 0.78
(-2.4, -2)	(0.608,0.643	(0.059, 0.121)	(0.00025, 0.0006)	(0.00039 0.00095)	(-227.272, -192.753)	(69.32,80.16)	(0.793, 0.821)
(-1.9, -1.5)	(0.651,0.682	(0.144, 0.285)	(0.0008, 0.0018)	(0.0012, 0.0028)	(-182.557, -151.024)	(83.08, 95.43)	(0.828, 0.85
(-1.4, -1)	(0.689,0.716	(0.336, 0.639)	(0.0023, 0.0058)	(0.0035 0.0087	(-144.143, -114.198)	(98.64, 111.7)	(0.863, 0.89
(-0.9, -0.5)	(0.721,0.741	(0.747, 1.346)	(0.0074, 0.019)	(0.011, 0.027)	(-108.287, -85.115)	(114.9, 127)	(0.896, 0.91
(-0.4, -0.1)	(0.745,0.755	(1.547, 2.289)	(0.024, 0.050)	(0.034, 0.069)	(-80.339, -66.168)	(129.7, 136.9)	(0.923, 0.936
0	0.758	2.585	0.064	0.088	-60.175	138.9	0.939
(0.1, 0.5)	(0.760,0.766	(2.906, 4.422)	(0.082, 0.219)	(0.112, 0.296)	(-54.741, -44.111)	(140.6, 145)	(0.943, 0.960)
(0.6, 0.9)	(0.766,0.767	(4.861, 6.292)	(0.282, 0.602)	(0.379, 0.798)	(-41.785, -38.111)	(145, 145.5)	(0.964, 0.972)
1	0.765	6.803	0.775	1.026	-37.370	144.4	0.973
(1.1, 1.5)	(0.757,0.763	(7.329, 9.569)	(0.999, 2.784)	(1.321, 3.671)	(-37.098, -36.431)	(138, 143.5	(0.975, 0.977)
(1.6, 2)	(0.742,0.754	(10.152, 12.520)	(3.601, 10.127)	(4.752, 13.465)	(-42.6777, -37.706)	(127.6, 136.	(0.975, 0.977)
(2.1, 2.5)	(0.722,0.739	(13.112, 15.777)	(13.128, 37.368)	(17.507, 50.373)	(-52.013, -44.152)	(115.5, 125.	(0.969, 0.974)
(2.6, 3)	(0.699,0.718	(16.514, 19.924)	(48.608, 140.006)	(65.708, 191.224)	(-63.463, -54.843)	(103.1, 113	(0.964, 0.969)

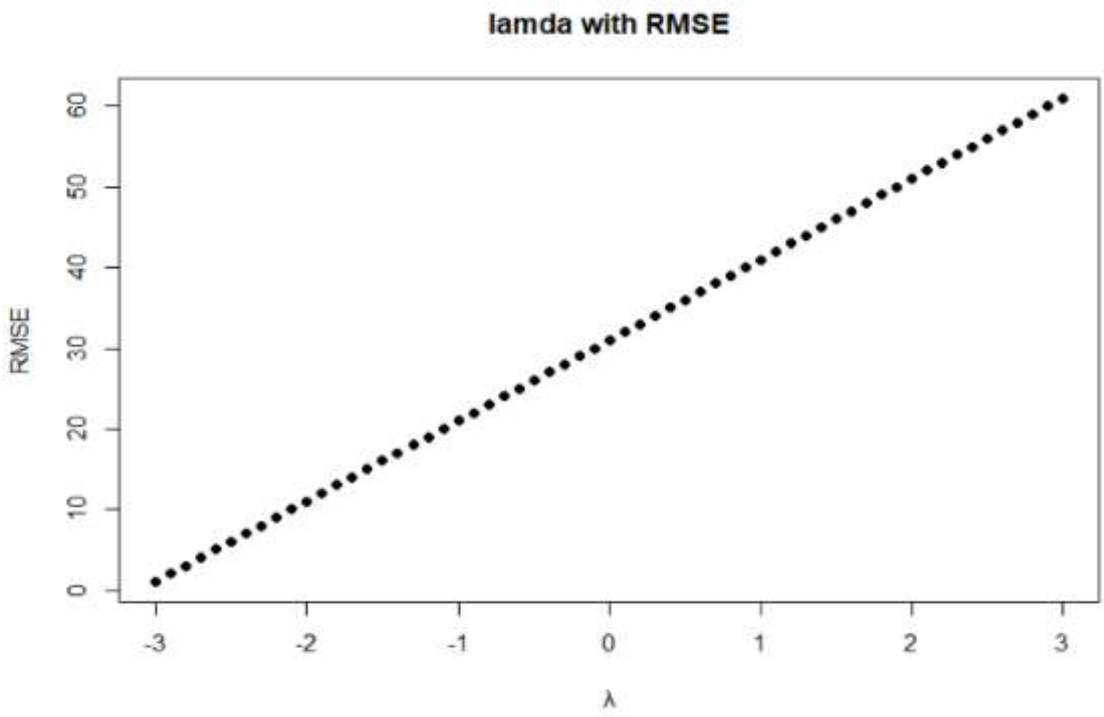
Table2: Selecting the best optimal power parameter according criteria

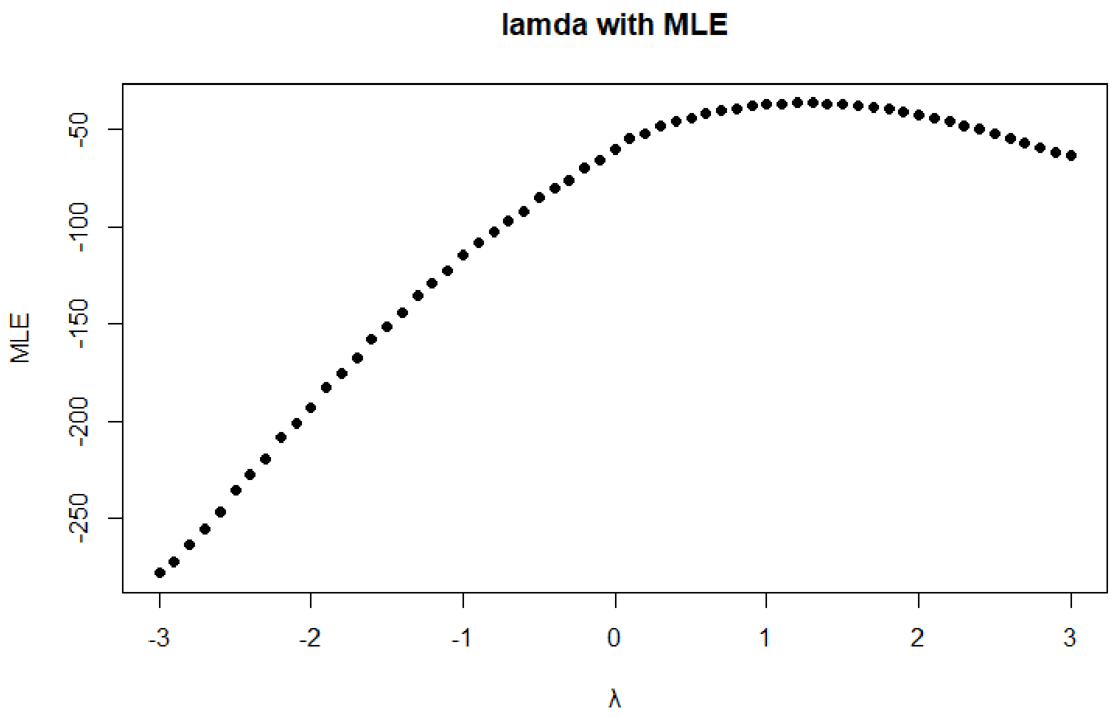
Power parameter	Criteria	Optimal value
0.7	$Adj.R^2$	0.7666
-3	*MAPE*	0.0202
-3	*MAE*	6.734E-05
-3	*RMSE*	0.0001
1.3	*MLE*	-36.4305
0.7	*F-statistic*	145.5
1.5	*SW*	0.9774




lamda with MAPE
MAPE
0
10
20
30
40
50
60
-3
-2
-1
0
1
2
3
λ


lamda with MAE
MAE
0
10
20
30
40
50
60
-3
-2
-1
0
1
2
3
λ


lamda with RMSE
RMSE
0
10
20
30
40
50
60
-3
-2
-1
0
1
2
3
λ


lamda with MLE
MLE
-50
-100
-150
-200
-250
-3
-2
-1
0
1
2
3
λ

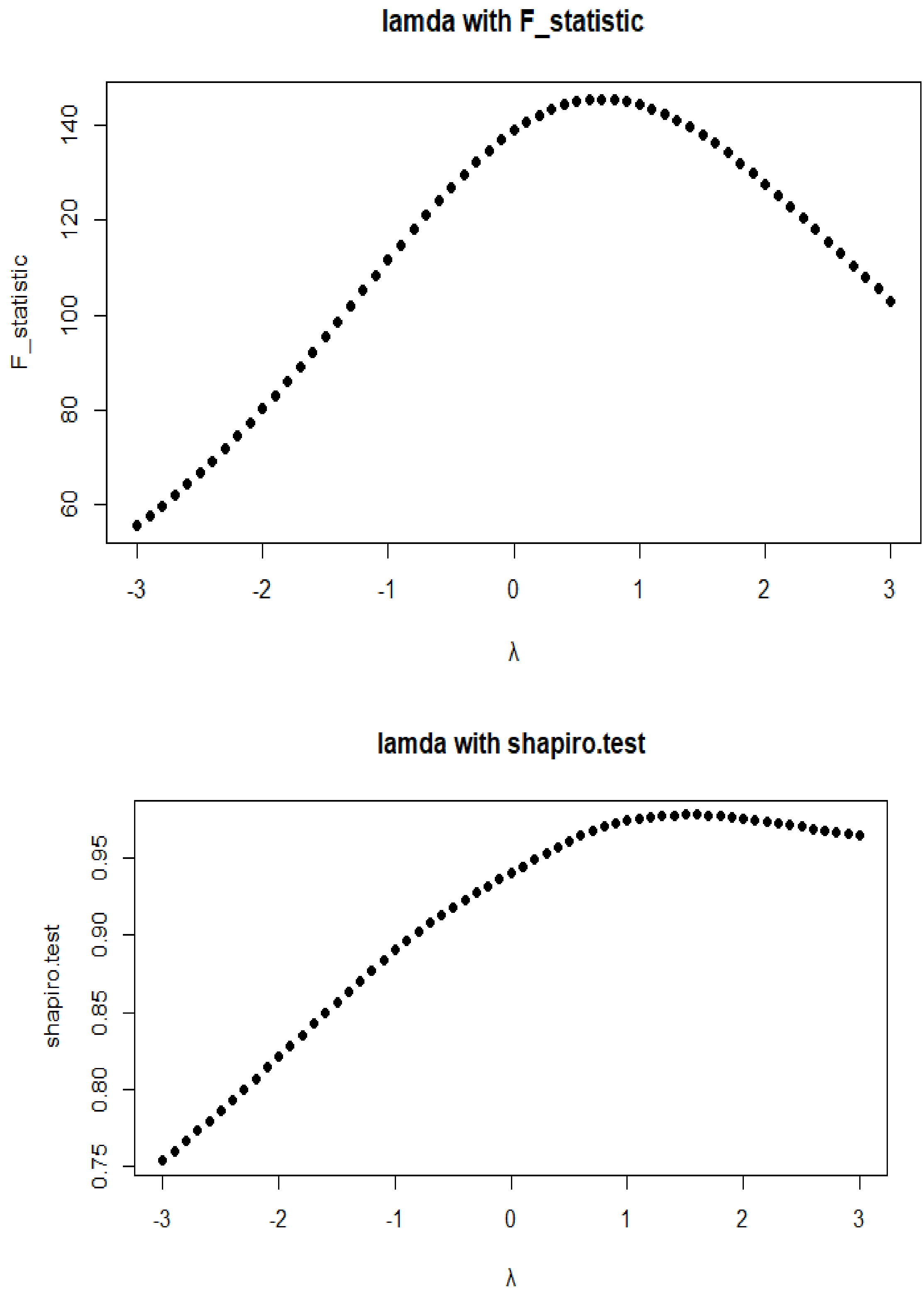


Figure 1. plot all statistical criteria against Lambda

In all plots, the horizontal axis represents the varying values of lambda, and the vertical axis illustrates how each evaluation criterion changes in response to these variations.

4. Conclusion

Power transformations play a crucial role in medical and health research, particularly for normalizing skewed data, stabilizing variance, and improving the validity of statistical models. While highly effective, they must be applied thoughtfully, with awareness of their impact on interpretability and inference. Emerging research seeks to refine their use in complex, high-dimensional biomedical data contexts. The primary objective of this research was to assess the effectiveness of the Box-Cox transformation within the general linear model (GLM) framework for analyzing factors affecting blood pressure. The study showed that when using the Box-Cox transformation with different lambda values, various evaluation criteria reacted differently to changes in lambda because each criterion emphasizes different model qualities, such as fit, residual normality, variance stabilization, or parsimony. Thus, identifying the "best" λ often depends on the primary goal of the analysis: maximizing likelihood, improving prediction, satisfying assumptions, or minimizing error. The consistency observed between criteria such as adjusted R-squared and the F-statistic indicates that stronger model performance was achieved when lambda was near 0.7. Adjusted R-squared reflects how well the model explains variance in the transformed response variable while penalizing for model complexity. However, a higher adjusted R-squared does not necessarily indicate better normality or variance stabilization. In contrast, the F-statistic indicates that the model explains a significant amount of variance relative to the error, suggesting that the predictors are jointly significant. Predictive error measures, such as MAPE, MAE, and RMSE, were minimized when lambda approached 3, showing an inverse relationship between lambda and prediction accuracy. The measures were prediction errors, preferring transformations that reduce variance without fully addressing distributional assumptions. Concurrently, maximum likelihood estimation (MLE) peaked around $\lambda = 1.3$, as it is directly influenced by how well the transformed data conform to the assumed normal distribution of residuals. This criterion typically reaches its maximum when the transformation effectively normalizes the data and stabilizes variance. The Shapiro–Wilk test for normality showed optimal results near $\lambda = 1.5$, specifically assessing the normality of residuals. These slight differences highlight how various criteria prioritize different aspects of model assumptions and fit.

Developing a statistical model that accurately reflects underlying data patterns is essential for producing reliable and interpretable results. The linear model remains a fundamental tool in statistical analysis due to its simplicity and broad applicability. Although it is often used as an initial model, transformations like Box-Cox are crucial when assumptions such as normality and homoscedasticity are violated. By applying the Box-Cox transformation within a general linear model (GLM) framework, we were able to enhance model robustness, reduce data distortions, and improve both interpretability and predictive accuracy.

This research adds valuable insight to the broader field of statistical modeling, especially in the health sector, where accurate modeling is directly linked to improving human well-being. Our study used clinical data related to blood pressure and analyzed the influence of various criteria. The findings emphasize the importance of integrating real-world health data with transformation techniques to produce more precise models.

After careful evaluation of multiple statistical performance criteria, we conclude that no single lambda value is optimal for all metrics. Instead, we recommend a range of lambda values that can be tailored to specific metrics, allowing for a more nuanced approach to model optimization. Selection should be based on the primary goal of the transformation, which is to improve model assumptions, thereby enhancing the reliability of parameter estimates and predictions. The transformation primarily ensures that the statistical assumptions underlying regression analysis are better satisfied. Using multiple evaluation criteria is essential to achieve a balanced and comprehensive model selection process.

The findings of this study may support health researchers and professionals in choosing appropriate modeling techniques for analyzing blood pressure and other health indicators. Future research will aim to explore additional transformation methods, incorporate new predictors, and apply the framework to broader clinical datasets, further enhancing the applicability and reliability of statistical models in healthcare research.

Acknowledgments
The authors are very grateful to the University of Duhok, and College of Administration and Economics, which helped improve this work's quality.

Conflict of Interest
The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Ethical Approval
Ethical approval was not required for this study as it did not involve human participants, personal data.

References

1. Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370–384. https://doi.org/10.2307/2344614
2. Dobson, A. J. (2002). An introduction to generalized linear models (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781315182780
3. Ruppert, D. (2001). Statistical Analysis, Special Problems of: Transformations of Data. International Encyclopedia of the Social & Behavioral Sciences, 15007- 15014. https://doi.org/10.1016/B0-08-043076-7/00513-1
4. Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. J R Stat Soc B Methodol, 26(2), 211–252. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x
5. Yeo, I., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87(4), 954–959. https://doi.org/10.1093/biomet/87.4.954
6. Yang, Z. (2006). A modified family of power transformations. Economics Letters, 92(1), 14–19. https://doi.org/10.1016/j.econlet.2006.01.011
7. 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 Research Methodology, 11(1), 118. https://doi.org/10.1186/1471-2288-11-118
8. Raymaekers, J., & Rousseeuw, P. J. (2020). Transforming variables to central normality. Machine Learning, 113(8), 4953–4975. https://doi.org/10.1007/s10994-021-05960-5
9. Hawkins, D. M. (2024). Testing normality of data transformed by maximum likelihood Box-Cox. School of Statistics, University of Minnesota. https://doi.org/10.48550/arXiv.2407.19329
10. Wichitaksorn, N., Choy, S. T. B., & Gerlach, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. Canadian Journal of Statistics, 42(4), 579–596. https://doi.org/10.1002/cjs.11228
11. Pek, J., Wong, O., & Wong, C. M. (2017). Data Transformations for Inference with Linear Regression: Clarifications and Recommendations. Practical Assessment, Research and Evaluation, 22(9), 1–11. https://doi.org/10.7275/2w3n-0f07
12. Marill, K. A. (2004). Advanced statistics: Linear regression, part II: Multiple linear regression. Academic Emergency Medicine, 11(1), 94–102. https://doi.org/10.1197/j.aem.2003.09.006
13. Lin, L., Jiang, W., Chen, B., Yu, J., & Zheng, C. (2024). Construction and Application of Cost Prediction Model Based on Multiple Linear Regression Analysis. In Procedia Computer Science (Vol. 247, Issue C, pp. 617–623). https://doi.org/10.1016/j.procs.2024.10.074
14. Ramachandran, K. M., & Tsokos, C. P. (2009). Mathematical statistics with applications. Burlington. Elsevier Academic Press.
15. Mohammed, A. H., & Mahdi, M. J. (2025). Comparison of estimation methods for the parameters of the Fréchet distribution using simulation. Iraqi Journal of Statistical Sciences, 22(1), 101–113. https://doi.org/10.33899/iqjoss.2025.187758
16. Hogg, R. V., McKean, J. W., & Craig, A. T. (2019). Introduction to mathematical statistics (8th ed.). Boston. Pearson. ISBN 978-0-13-468699-8.
17. Atkinson, A. C., Riani, M., & Corbellini, A. (2021). The Box–Cox Transformation: Review and Extensions. Statistical Science, 36(2), 239–255. DOI: 10.1214/20-STS778
18. Wohlwend, B. (2023). Regression model evaluation metrics: R-squared, adjusted R-squared, MSE, RMSE, and MAE. Medium.
19. Pirenne, S., & Claeskens, G. (2024). Exact post-selection inference for adjusted R squared selection. Statistics and Probability Letters, 211, 11013. https://doi.org/10.1016/j.spl.2024.110133
20. Miles, J. (2005). R Squared, Adjusted R Squared. Wiley StatsRef: Statistics Reference Online, 4, 1655–1657. https://doi.org/10.1002/0470013192.bsa526
21. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014
22. de Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. Neurocomputing, 192, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114
23. Gupta, S. C., & Kapoor, V. K. (2000). Fundamentals of mathematical statistics: A modern approach (10th rev. ed.). Sultan Chand & Sons.
24. Seber, G. A. F., & Lee, A. J. (2003). Linear regression analysis (2nd ed.). Hoboken. John Wiley & Sons. https://doi.org/10.1002/9780471722199.ch1
25. El-horbaty, Y. S. (2024). A Note on Effective Transformation-based Exact F-test for Sub- Clustering Effect in Two-Fold Nested Error ANOVA Model. Journal of Statistics and Computer Science, 3(1), 79-89. https://doi.org/10.47509/JSCS.2024.v03i01.05
26. Khatun, N. (2021). Applications of Normality Test in Statistical Analysis. Open Journal of Statistics, 11(01), 113–122. DOI:10.4236/ojs.2021.111006

تحليل تأثير معايير مختلفة على ضغط الدم من خلال نموذج خطي عام باستخدام تحويلات بوكس–كوكس

زيبا محمد مصطفى[1] ازاد عادل شريف[2]

[2,1] قسم الاحصاء والمعلوماتية, كلية ادارة وافتصاد, جامعة دهوك, اقليم كوردستان, عراق.

الخلاصة: ضغط الدم هو مؤشر حيوي لصحة القلب والأوعية الدموية، حيث يُمثل القوة التي يمارسها الدم المتدفق على جدران الشرايين. تحويل البيانات تلعب دورًا مهمًا في تلبية الافتراضات الأساسية المطلوبة للنماذج الخطية في النمذجة الإحصائية ، مثل التوزيع الطبيعي، الخطية، وتجانس التباين، وخاصةً عند عدم وجود هذه الافتراضات. تساعد هذه التحويلات على تصحيح الاخطاء في هيكل البيانات، مما يحسن من صحة وموثوقية النماذج الخطية العام(GLMs). في هذه الدراسة، استخدمنا تطبيق تحويل بوكس–كوكس(Box–Cox) ، وهو أحد تقنيات تحويلات القوى والتي يهدف تحسين أداء نموذج الانحدار الخطي. والهدف الرئيسي للدراسة تحديد أفضل قيمة لمعامل التحويل (λ) من أجل تحسين ملاءمة النموذج وقابليته للتفسير، في هذه الدراسة تم استخدام بيانات تحليل الدم التي جمعت من مستشفى آزادي في محافظة دهوك .طبقت عدة معايير إحصائية لاختيار افضل قيمة لمعامل التحويل، بما في ذلك: معامل التحديد المعدل(Adj. R^2) ، متوسط الخطأ المطلق(MAE) ، متوسط النسبة المئوية للخطأ المطلق(MAPE) ، قيمة الإحصائيةF ، تقدير الاحتمالية العظمى(MLE) ، الجذر التربيعي لمتوسط مربع الخطأ (RMSE)، واختبار شابيرو–ويلك .(Shapiro–Wilk) وقد استُخدمت هذه المعايير لتقييم جودة النموذج وللمساعدة في اختيار افضل نموذج الانحدار الخطي العام. وتشير نتائج البحث إلى أن القيم المختلفة لمعامل λ تحسن جوانب مختلفة من أداء النموذج حسب المعايير التي طبقت في الدراسة و نستنتج ذلك عندما $\lambda = 0.7$ يحقق توازنًا بين جودة ملاءمة النموذج وقوته التفسيرية (ارتفاع Adj. R^2 ، وقيمة مرتفعة مناسبة لإحصائية.F), $\lambda = -3$ يقدم أكثر التنبؤات دقة مع أدنى معدلات للخطأ و $\lambda = 1.3$ او 1.5 يعزز تقدير الاحتمالية ويحسن التوزيع الطبيعي للبواقي. ولتعزيز هذه العملية، اقترحنا خوارزمية حسابية لتقدير القيمة المثلى لـ λ. وبناءً على معيار معامل التحديد المصحح، حددت الدراسة قيمة مثالية لمعامل التحويل أدت إلى نموذج قوي وذو ملاءمة عالية

الكلمات المفتاحية : تحويل بوكس–كوكس، المعايير الإحصائية، النماذج الخطية العامة واختبار شابيرو–ويلك