Experimenting The Box-Cox Family Transformation on Likert Scale Data for Non-Normal Residuals in Linear Regression

Mohd Azry Abdul Malik[1]*, Nur Alia Sofea Ahmad Fauzi[2], Nor Aina Nabila Japri[3], Nur Khaliesah Supardi[4], Amri Ab Rahman[5], Omar Kairan[6], Jasrul Nizam Ghazali[7], Muhammad Firdaus Mustapha[8]

[1,2,3,4,5,6,8]*Faculty of Computer & Mathematical Science, Universiti Teknologi MARA Kelantan, Malaysia.*
[7]*Centre of Foundation Studies, Universiti Teknologi MARA Selangor, Dengkil Campuses, 43800 Dengkil, Selangor, Malaysia.*
**Email: azry056@uitm.edu.my*

Abstract

Neglecting the necessity for normality of residuals leads to failing in meeting the assumptions of error term in linear regression. Addressing a violation of this assumption requires an appropriate data transformation. The first objective of this study is to identify a suitable Box-Cox transformation (BCT) family for Likert scale data to handle non-normal residuals in multiple linear regression (MLR). When confronting with non-normally distributed MLR residuals, some scholars argue that the ordinary least squares estimation approach, commonly used in linear regression, consistently produces a reliable estimated value even when the error term deviates from a normal distribution. Given these conflicting opinions, one asserting that non-normally distributed error terms result in inaccurate estimates and the other maintains that such deviations do not compromise the consistency of estimates; thus, the second objective of this experiment is to reaffirm the differences in viewpoint by comparing the consistency of estimation values in MLR between cases of normal (transformed) and non-normal (non-transformed) residuals. The study suggested that the optimal BCT occurred at a lambda value 0.5. This specific lambda value corresponds to a logarithmic transformation, signifying a fundamental shift in Likert scale data toward a more normalized distribution or residuals. In the context of conflicting opinions regarding the impact of non-normally distributed error terms on estimates in MLR, this study revealed a significant difference in the mean of estimated values between the transformed and non-transformed models. The empirical evidence suggests that non-normally distributed error terms do lead to inconsistency in estimation values in MLR. Appropriate transformation does contribute to more reliable and interpretable results in MLR.

Keywords: Box-Co; Transformation; Likert scale; Linear regression

1. Introduction

Regression analysis is a statistical method used to analyze the relationship between quantitative variables, enabling predictions of a dependent variable based on independent variables. Linear regression, categorized into simple linear regression (SLR) and multiple linear regression (MLR), is valid only when four key assumptions are met: linearity, independence, homoscedasticity, and normality of residuals. When these assumptions are violated, data transformations become necessary to align the data with these requirements. Transformations improve data normality, equalize variance, and prevent the exclusion of observations, making the dataset more suitable for analysis (Knief and Forstmeier, 2021).

The Box-Cox transformation (BCT) is a widely used method for addressing assumption violations in linear regression. It includes logarithmic, square root, and reciprocal transformations as specific cases. BCT optimizes data normality through a lambda

parameter (λ), which determines the transformation's extent. The optimal λ value is determine by maximizing the log-likelihood function or minimizing residual sum of squares. This approach is particularly useful for correcting issues like heteroscedasticity or non-normal errors, thus enhancing the reliability and interpretability of regression models.

Likert scale data, commonly used in social sciences to measure subjective responses like attitudes or perceptions, presents challenges for linear regression due to its ordinal nature. Although Likert scales approximate interval-level measurements when well-constructed and symmetric (Hair et al., 2017), they often fail to meet normality assumptions. Transformations like BCT can make Likert scale data suitable for MLR by addressing non-normal residuals while maintaining statistical rigor. The first objective of the study aims to identify to identify a suitable BCT family method for handling such deviations of non-normal residuals in Likert scale data.

Second, misconceptions about linear regression often involve misunderstandings about error term normality. While some argue that ordinary least squares (OLS) estimators remain consistent despite non-normal errors (Copeland, 1997), others emphasize that such violations can lead to inaccuracies in hypothesis testing and confidence intervals (Schmidt and Finan, 2018). By comparing estimation consistency between transformed (normal) and non-transformed (non-normal) residuals in MLR, this study seeks to clarify these conflicting viewpoints as the second objectives and contribute to a better understanding of how transformations impact regression analysis results.

2. Literature Review

In statistical analysis, ensuring that residuals are normally distributed is crucial for accurate results. A common method to check this assumption is by using a Q-Q plot, where deviations from a straight line suggest that the data is not normally distributed. Non-normal data can lead to biased results, inefficient estimates, and incorrect model evaluations (Malik, 2018). Therefore, researchers need to address this issue using either non-parametric tests or data transformations (Sachin and Us, 2019).

Non-parametric tests provide an alternative when data doesn't meet normality assumptions. Examples of these tests include the Wilcoxon rank sum test, the Mann-Whitney test, the Moods Median test, and the Kruskal-Wallis test (Sachin and Us, 2019). These tests are particularly useful for nominal or ordinal data and for testing hypotheses that do not involve specific population parameters. However, they are generally less sensitive and efficient compared to parametric tests when the assumptions for parametric tests are met (Mitek, 2022; Pek et al., 2018).

Data transformations offer another approach to handle non-normal data by making the dataset more closely resemble a normal distribution (Pek et al., 2018). Common transformations include log, square root, and arcsine transformations, which are also known as monotonic transformations because they apply a mathematical function independently to each data value, preserving the original data's rank.

Likert scales, which are widely used in survey research, provide a standardized way for respondents to express their opinions or agreement with various statements (Pimentel, 2019). Researchers often use odd-numbered scales, such as five-, seven-, or nine-point scales, to capture intermediate responses (Taherdoost, 2019). However, when using Likert scales,

researchers need to be aware of potential response biases, like acquiescence bias and extreme response bias, which can affect the validity of the data (Suárez-Alvarez et al., 2018). Additionally, the limited range of response options in Likert scales may not fully represent the complexity of respondents' views, and the resulting data often do not follow a normal distribution, which violates the assumptions of many parametric tests.

To address these issues, researchers can use the Box-Cox transformation (BCT), a statistical technique developed by George Box and David Cox. The BCT is used to transform non-normal dependent variables into a normal shape (Malik, 2018) and is valuable when fundamental assumptions of a regression model are violated. It offers a family of transformations, including square root, cube root, natural log, and inverse transformations, to optimally normalize the data, increase correlation coefficients, and improve the accuracy of statistical analyses (Malik, 2018).

3. Methodology and Data Collection

Ethics Approval

The Ethical approval for this study has been provided by the UiTM Research Ethics Committee, and the associated reference number is 500-CK (PJIA/URMI 5/1/1).

Research Design

This study investigates methods for handling non-normal residuals in statistical analyses, particularly when using Likert scales. Likert scales, frequently used in social science research, provide standardized response options but often produce data that violates normality assumptions. To examine this issue, the study employed a simulation approach, using R to generate data mimicking a 9-point Likert scale (Hair et al., 2017; Malik et al., 2021) with intentionally non-normal residuals. The research had two primary objectives: first, to identify a suitable Box-Cox transformation (BCT) to normalize the residuals in multiple linear regression (MLR) involving the simulated Likert scale data; and second, to compare the consistency of estimation values between MLR models using the data with and without the BCT applied. By applying the BCT with different lambda values and assessing normality via Q-Q plots, the study aimed to recommend best practices for BCT usage with Likert scales. Applying the BCT helps to determine the optimal transformation for normalizing the required study variable (Aufa et al., 2018). Next, the impact of non-normality on estimation was then evaluated using a repeated measures t-test. Figure 1 show the process flow for the study.

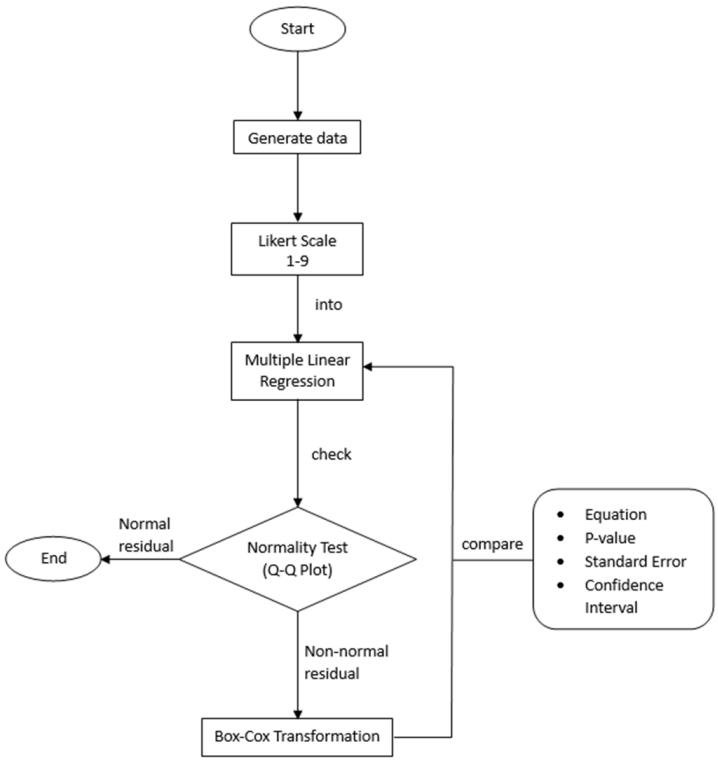


Figure 1: Process flow

Data Generating Mechanism

The data simulation is performed using R programming. The details command is as in Table 1.

Table 1: *Command for data generating*

Description	Command
Install additional distributions	Install.packages ("extraDistr")
Using the installed package	library(extraDistr)
Generate dependent variable	variable1 <- rpois(1000, 3) y <- sample(variable[variable >= 1 & variable <= 9], 150, replace=TRUE)
Generate independent variables	variable2 <- rpois(1000, 3) x1 <- sample(variable[variable >= 1 & variable <= 9], 150, replace=TRUE) x2 <- sample(variable[variable >= 1 & variable <= 9], 150, replace=TRUE) x3 <- sample(variable[variable >= 1 & variable <= 9], 150, replace=TRUE)

Normality checking

The Kolmogorov-Smirnov test is used to check if the residuals are normally distributed. This test is suitable for larger sample sizes (n≥50), while the Shapiro-Wilk test is more appropriate for smaller samples (<50) (Mishra et al., 2019). The details command is as in Table 2.

Table 2: Command for MLR and normality plot/test

Command	Description
model =lm(y~x1+x2+x3)	Create a linear model
qqnorm(model $residuals) qqline(model $residuals)	Create a Q-Q plot
ks.test(model $residuals)	Test of normality

Applying the Transformation

This study aims to identify an appropriate BCT family for Likert scale data to address non-normal residuals in MLR. The BCT involves finding an optimal lambda value between -3 and +3 to transform the dependent variable into a normal distribution (see Table 3 for common transformations). The optimal lambda is chosen based on how closely it approximates a normal distribution curve. The R commands for this transformation are in Table 4.

Table 3*:* Common Box-Cox transformations

Lambda value	Transformed data	Lambda value	Transformed data
-3	$Y^{-3} = \frac{1}{Y^3}$	0.5	$Y^{0.5} = \sqrt{Y}$
-2	$Y^{-2} = \frac{1}{Y^2}$	1	Y
-1	$Y^{-1} = \frac{1}{Y^1}$	2	Y^2
-0.5	$Y^{-0.5} = \frac{1}{\sqrt{Y}}$	3	Y^3
0	$\log Y$		

Table 4: Command for Box-Cox family transformation

Description	Command
Install package	Install.packages("MASS")
Using the installed package	library(MASS)
Specify lambda	lambda <- 0.5
Transform the dependent variable	transformed_y <- if (lambda != 0) (y^lambda - 1) / lambda else log(y)

Comparing the Estimation Value

To address conflicting views on the impact of non-normal residuals on linear regression estimates, this study used a repeated measures t-test. This test compares the mean estimation values from multiple linear regression (MLR) models with normal (transformed) and non-normal (non-transformed) residuals, assessing whether the transformation significantly alters the model's estimations. R commands for the repeated measures t-test are provided in Table 5.

Table 5: Command for repeated measures T-test

Description	**Command**
Create data predicted value	data$predicted=predict(model) datasettransform$predicted <- predict(model_transformed)
Compare model using paired t-test	t.test(data$predicted, datasettransform$squared_predicted,paired=T)

4. Results

Box-Cox Transformation on Non-normal Residuals

The first objective of this study is to identify an appropriate Box-Cox transformation (BCT) family for Likert scale data, specifically to address non-normal residuals in Multiple Linear Regression (MLR). Various Box-Cox transformation families, using different lambda values, have been tested on Likert scale data to address the issue of non-normal residuals in the

context of multiple linear regression. The suitability of lambda values for normalizing residuals through BCT is determined using the Kolmogorov-Smirnov test and Q-Q plot.

Table 6 displays the results of the Kolmogorov-Smirnov test for the simulated data, when no transformation performed (λ=1). While Table 7 presents the results of the Kolmogorov-Smirnov test, where the D-statistic measures the maximum difference between the cumulative distribution functions of two datasets. This test is employed to compare the distribution of residuals to a theoretical normal distribution. A lower D-statistic and a higher p-value are favorable, indicating that observed residuals closely align with the assumed normal distribution. Conversely, a higher D-statistic and a lower p-value suggest a significant disparity between observed and expected distributions. A large D-statistic implies rejection of normality, while large p-values indicate normally distributed data.

Among all the models compared, Model 6 (λ=0.5) is identified as the optimal transformed model. In Model 6, the D-statistic value is 0.044166, which is smaller than the significance level (α=0.05) and is the smallest among all models. Additionally, it boasts the highest p-value of 0.9316. These findings provide strong evidence supporting the notion that residuals in Model 6 follow a normal distribution. The Q-Q plot for Model 6 further reinforces this conclusion by displaying points along a straight line, indicating a normal distribution for the obtained residuals.

Table 6: Kolmogorov-Smirnov normality test and Q-Q plot on non-transformed model

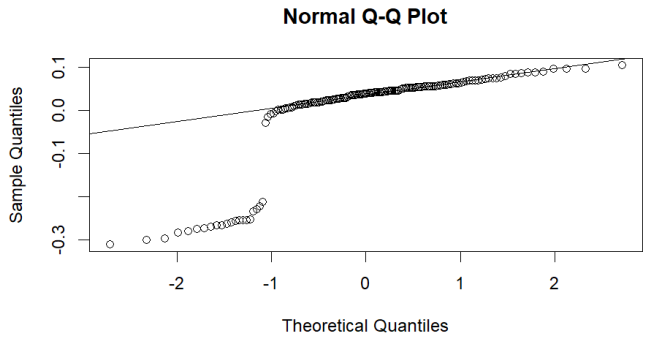
Kolmogorov-Smirnov test		Normal Q-Q Plot
D-statistic	P-value	
0.1476	0.002902	

Table 7: Kolmogorov-Smirnov normality test and Q-Q plot on BCT

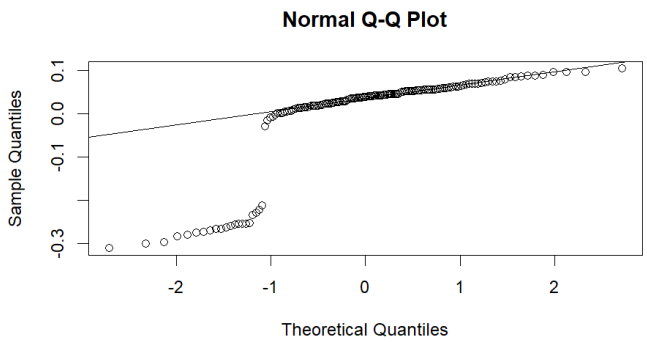
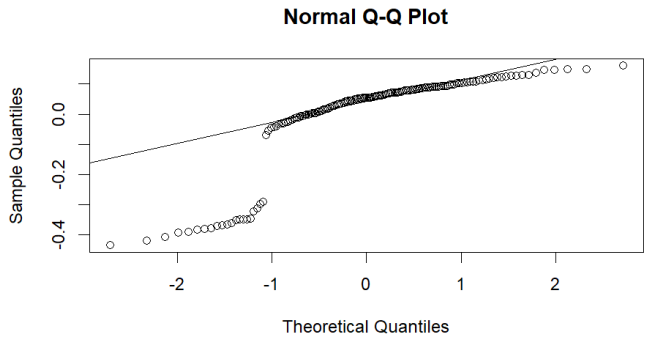
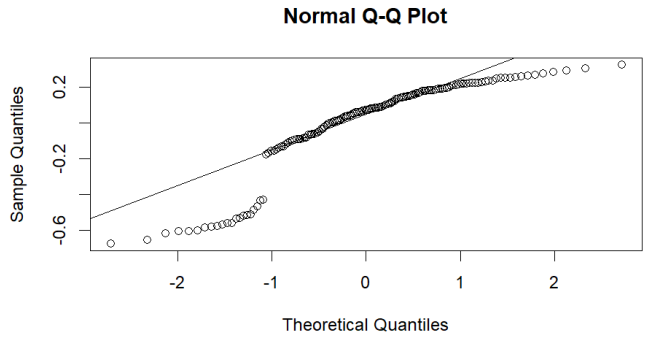
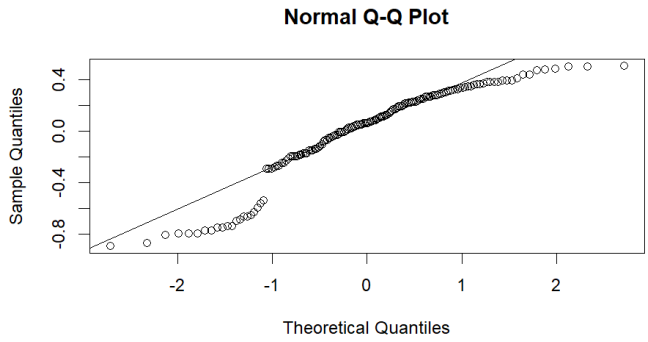
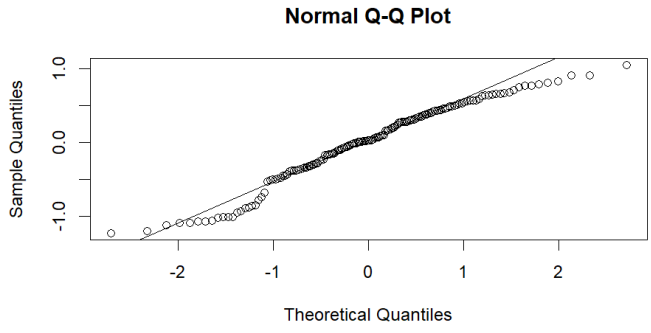
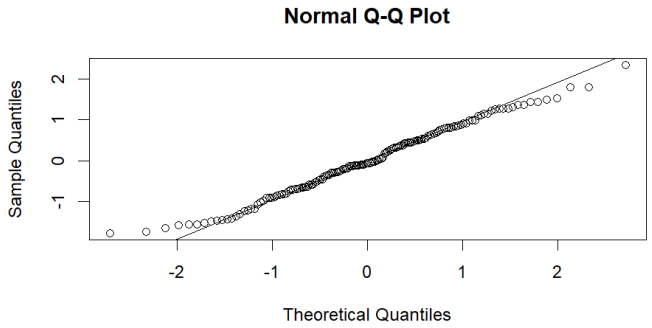
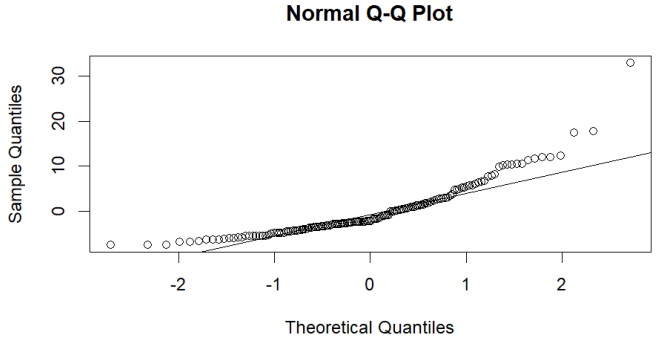
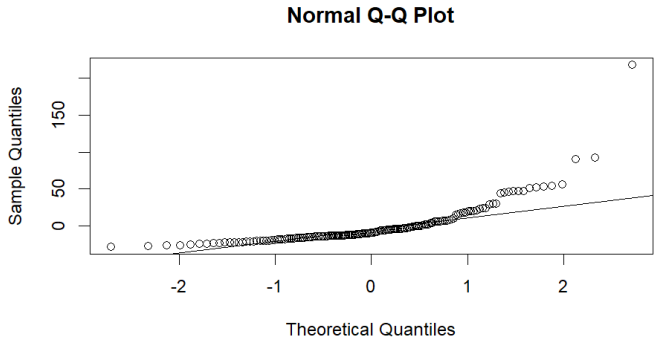
Model	Lambda value	Kolmogorov-Smirnov test		Normal Q-Q Plot
		D-statistic	P-value	
1	-3	0.45858	< 2.2e-16	

Table 7 (Continue): Kolmogorov-Smirnov normality test and Q-Q plot on BCT

Model	Lambda value	Kolmogorov-Smirnov test		Normal Q-Q Plot
		D-statistic	P-value	
2	-2	0.43601	< 2.2e-16	
3	-1	0.37225	< 2.2e-16	
4	-0.5	0.30686	1.079e-12	
5	0	0.18368	8.045e-5	

Model	Lambda value	Kolmogorov-Smirnov test		Normal Q-Q Plot
		D-statistic	P-value	
6	***0.5***	***0.044166***	***0.9316***	
7	2	0.49196	< 2.2e-16	
8	3	0.65103	< 2.2e-16	

Consistency of Estimation Values in MLR Between Normal and Non-normal Residuals Cases

The second objective of this experiment is to assess the consistency of estimation values in Multiple Linear Regression (MLR) when comparing cases with normal (transformed) and non-normal (non-transformed) residuals. The repeated measures t-test is a reliable method for comparing the mean difference in estimated values between the transformed and non-transformed models. According to Table 8, the p-value being less than the significance level (α=0.05) provides sufficient evidence to conclude a significant difference in the mean of estimated values between the transformed and non-transformed models. Tables 9 and 10 present the regression coefficients for the transformed (Model 6) and non-transformed models, respectively. Additionally, Table 11 displays the standard error and Analysis of Variance (ANOVA) values for both models. The transformed model exhibits a smaller

standard error value of 0.882 compared to the non-transformed model (1.558). Moreover, the transformed model has a slightly larger ANOVA value than the non-transformed model, thereby increasing the probability (lower p-value) of being considered a significant model.

Table 8: Result of repeated measure T-test

T	P-value	95 % Confidence Interval		Mean difference
		Lower	Upper	
55.659	< 0.00000000000000022	0.982451	1.054777	1.018614

Table 9: Coefficients of non-transform model

Parameter	Estimate	Standard Error	T-value	Sign	Confidence Interval	
					Lower	Upper
Intercept	3.142108	0.418085	7.515	0.0000	2.3158	3.9684
x_1	-0.002058	0.076846	-0.027	0.9787	-0.1539	0.1498
x_2	-0.132174	0.080415	-1.644	0.1024	-0.2911	0.0268
x_3	-0.171159	0.076837	-2.228	0.0274	-0.3230	-0.0193

Table 10: *Coefficients of transform model*

Parameter	Estimate	Standard Error	T-value	Sign	Confidence Interval	
					Lower	Upper
Intercept	1.977671	0.236697	8.355	0.0000	1.5099	2.4455
x_1	0.009095	0.043506	0.209	0.8347	-0.0769	0.0951
x_2	-0.080215	0.045527	-1.762	0.0802	-0.1702	0.0098
x_3	-0.096096	0.043501	-2.209	0.0287	-0.1821	-0.0101

Table 11: *Significant model between best transformer and non-transform model*

Statistic	Model	
	Transform	Non-transform
Lambda value	0.5	1
Standard Error	0.882	1.558
F-test	2.701	2.621
P-value	0.04787	0.05300

5. Conclusions and Recommendations

This study aimed to address issues with non-normal residuals in multiple linear regression (MLR) using Likert scale data, with two primary objectives: identifying the best Box-Cox transformation (BCT) family and evaluating the consistency of estimation values in MLR with and without data transformation. Analysis revealed that a lambda value of 0.5, corresponding to a square root transformation, was optimal for normalizing the Likert scale data, supported by low D-statistics and a p-value greater than 0.05. This suggests that using a square root transformation with lambda 0.5 is an effective way to improve residual normality and enable more robust MLR studies.

The study found a significant difference in the mean of estimated values between transformed and non-transformed models, indicating that non-normally distributed error terms in MLR do lead to inconsistency in estimation values. The empirical evidence demonstrates that appropriate transformation contributes to more reliable and interpretable results in MLR.

To improve the study's practical application, it's recommended to test various transformation methods beyond just the Box-Cox transformation and to incorporate real-world Likert scale datasets alongside simulation data. By experimenting with different techniques and parameters and comparing results, researchers can gain a more nuanced understanding of how transformations impact residual normality. Furthermore, the utilization of real data not only confirms the transformation's effectiveness but also assures that the conclusions are directly applicable to realistic research situations.

References

Aufa, N., Ishak, M., & Ahmad, S. (2018). Estimating optimal parameter of Box-Cox transformation in multiple regression with non-normal data. *Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016) (Issue Rcstss)*. Springer Singapore. https://doi.org/10.1007/978-981-13-0074-5

Copeland, K. A. F. (1997). Applied Linear Statistical Models. In *Journal of Quality Technology*, *29*(2). https://doi.org/10.1080/00224065.1997.11979760

Hair Jr, J. F., Matthews, L. M., Matthews, R. L., & Sarstedt, M. (2017). PLS-SEM or CB-SEM: Updated guidelines on which method to use. *International Journal of Multivariate Data Analysis, 1*(2), 107-123.

Knief, U., & Forstmeier, W. (2021). Violating the Normality Assumption may be the Lesser of Two Evils. *Behavior Research Methods, 53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Malik, M. A. A., Mustapha, M. F., Sobri, N. M., Abd Razak, N. F., Zaidi, M. N. M., Shukri, A. A., & Sham, M. A. L. Z. (2021). Optimal reliability and validity of measurement model in confirmatory factor analysis: Different likert point scale experiment. *Journal of Contemporary Issues and Thought, 11,* 105-112.

Malik, F. (2018). Box-Cox transformation approach for data normalization: A study of new product development in manufacturing sector of Pakistan. *IBT Journal of Business Studies, 14*(1), 110–119. https://doi.org/10.46745/ilma.jbs.2018.14.01.09

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia, 22*(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

Mitek. (2022). Advantages and disadvantages of non-parametric tests. *UKEssays,* 26–29. https://www.ukessays.com/essays/communications/interpersonal-relationships-advantages-2866.php

Pek, J., Wong, O., & Wong, A. C. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in psychology*, *9*, 2104. https://doi.org/10.3389/fpsyg.2018.02104

Pimentel, J. L. (2019). Some biases in likert scaling usage and its correction. *International Journal of Sciences: Basic and Applied Research, 45*(1), 183–191.

Sachin, A., & Us, C. (2019). Non-normal data: How to deal with it? https://shorturl.at/efIU4

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology, 98*(0), 146–151. https://doi.org/10.1016/j.jclinepi.2017.12.006

Suárez-Alvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in likert scales: A questionable practice. *Psicothema, 30*(2), 149–158. https://doi.org/10.7334/psicothema2018.33

Taherdoost, H. (2019). What is the best response scale for survey and questionnaire design; review of different lengths of rating scale/attitude scale/Likert scale. *International Journal of Academic Research in Management (IJARM)*, *8*(1), 2296–1747.