Scholars Journal of Physics, Mathematics and Statistics

OPEN ACCESS

Abbreviated Key Title: Sch J Phys Math Stat
ISSN 2393-8056 (Print) | ISSN 2393-8064 (Online)
Journal homepage: https://saspublishers.com

Optimal Inverse Square Root Transformation for Response Surface Methodology: A Simulation Approach

Kupolusi, Joseph A[1], Mumini Saheed Lekan[1*]

[1]Department of Statistics, Federal University of Technology Akure, Nigeria

DOI: https://doi.org/10.36347/sjpms.2025.v12i03.005 | **Received:** 30.03.2024 | **Accepted:** 09.05.2024 | **Published:** 27.03.2025

***Corresponding author:** Mumini Saheed Lekan
Department of Statistics, Federal University of Technology Akure, Nigeria

Abstract | **Original Research Article**

Transformation techniques in Response Surface Methodology has received little attention in the literature when the response of design variables are not normally distributed. Various attempts to mitigate the incessant reoccurrence of the violation of normality assumption has proved abortive. This research delve into intricate of proposing inverse square root transformation, a robust transformation technique that can handle both small and large sample sizes in response surface methodology paradigm. Some transformation techniques in literature used for RSM are log, Box-Cox, square-root. These were tested and compared alongside with a newly proposed method. A Monte Carlo Simulation of different sample sizes (n = 10, 20, 50, 100, 200, 500, 1000) at different initial guess parameter for both small and large samples were used. Three tests Anderson-Darlington, Shapiro-wilk and Jarque-Berra test statistics were used to validate the consistency of the transformation methods considered and graphical representation were also used for visual inspection of the behavior of the methods. The result of the analysis revealed that inverse square root transformation method outperformed other existing method. This is achieved through comparison analysis of the methods using Bayesian Information Criterion (BIC).

Keywords: Normality test, QQ-plot, Monte Carlo Simulation, Assumption of Normality, Sample Size, Classical Tests.

1.0 INTRODUCTION

The statistical techniques are supported by a number of underlying assumptions. The assumption of normalcy is one of them. In many data studies, even when normalcy is assumed implicitly or conveniently, it is often necessary to verify it. In the event that the assumption is not met, the models' interpretations and conclusions are not valid, if not trustworthy. There are two methods for determining normalcy. The contrasts between the theoretical distribution, which is similar to a normal distribution, and the empirical distribution are visualized using graphical approaches. The statistical analyses conducted to test the null hypothesis that the variable has a normal distribution are carried out by the numerical methods. Plots are used by the graphical approaches to show the distribution. They are divided into theoretical and descriptive categories. Whereas the latter takes into account both theoretical and an empirical distribution, the former approach is grounded in empirical data. In data analysis, a transformation is the substitution of a variable for a function of that variable, as substituting the square root or logarithm of x for x. This alteration can significantly change the distribution or relationship represented by the data.

When using parametric analysis to determine whether or not the population in this case is normal, it is assumed that the population is normal. Normalcy tests are used in many sectors. One context for using normality testing is the residuals of a linear regression model. If the residuals are not normally distributed, they should not be used in Z tests or any other tests that are based on the normal distribution, such as t tests, F tests, and chi-squared tests. It's possible that the dependent variable, or at least one explanatory variable, has the erroneous functional form and that there are no significant variables, etc., if the residuals are not normally distributed.

This study will therefore concentrate on using various sample sizes to transform data to normalcy using various statistical tests and also using graphical plot to establish normality assumptions due to non-normality assumptions.

The assumption that errors (model residuals) are normally distributed is one of the most well-known presumptions in parametric statistics (Lumley *et al.*, 2002). The most extensively used tests for statistical significance, namely linear models (lm) and linear mixed models (lmm) with Gaussian error (which include the frequently more well-known procedures of regression, t test, and ANOVA), are based on this "normality assumption." Empirical data, on the other hand, frequently deviates significantly from normalcy and can even be categorical, like count or binomial data. The authors in [7], said that normalizing data to a normal distribution is an essential step that is made easier by normalization. It can be difficult to select the best transformation because different methods are more effective with different kinds of data. By automatically choosing the optimal transformation from a list of choices based on how effectively it normalizes the data, this tool provides a solution. Additionally, it makes it simple to integrate data pre-processing with other machine learning technologies. If necessary, you can even specify your own unique transforms. If one or more of these systematic mistakes are corrected, normally distributed residuals might result. In order to assess right-skewed data with zeros and negatives, the author in [1], uses inverse hyperbolic sine (IHS) transformation—which is frequently employed in economics. The main conclusion is that the modified data's units of measurement have a big influence on the outcomes, which could lead to bad business or policy judgments. The researchers suggest a technique for choosing suitable units for IHS-transformed data in order to overcome this. In addition to transformation methods, Real-world data frequently deviates from basic data assumptions, such as normality and constant variance, which are the foundation of parametric statistical approaches. In order to solve this problem, the author in [3], introduces variable transformation which is discussed in this study. Researchers can prevent statistical errors and prepare their data for parametric analysis by converting variables. Achieving unambiguous interpretations of the results is the ultimate goal, but it's crucial to keep in mind that transformations change the data's original units and meaning.

1.1 Descriptive plots and Theoretical Plots

While histograms are a popular tool for evaluating data normality, statisticians also use P-P and Q-Q plots for a more accurate assessment. These graphs contrast the distribution of the observed data with a theoretical distribution, usually the normal distribution. The Q-Q plot examines the actual data quintiles, whereas the P-P plot concentrates on the cumulative probabilities of the data. If the data is regularly distributed, the plot will be a straight line in both scenarios. It's critical to keep in mind that these are merely visual aids and cannot offer a conclusive response regarding normalcy. All they do is show how well the data fit the normal distribution.

1.2 Theoretical Statistics

To determine whether data has a normal distribution, there are various statistical tests available. The Kolmogorov-Smirnov (K-S) (smirnov, 1948) test and the Shapiro-Wilk test are two often utilized alternatives. To compare the observed data to a theoretical normal distribution, these tests—as well as the Cramer-von Mises (SAS institute 1995, von Mises, 1928). And Anderson-Darling tests (Anderson & Darling, 1954), all rely on the idea of the empirical distribution function (EDF). The Shapiro-Wilk, Jarque-Bera, and Anderson-Darling tests are particularly used in this study to assess data normalcy.

Table 1: Numerical tests of normality

Test	Statistic Sample Size (N)	Distn
Shapiro test	W	$\chi^2(2)$
Jarque-berra test	χ^2	$\chi^2(2)$
Anderson- Darlington	A^2	EDF

2.0 METHODOLOGY

A statistical technique called the Shapiro-Wilk test is used to determine how much a dataset resembles a normal distribution. The variance of the data is compared to the predicted variance of a perfectly normal distribution with the same mean in order to achieve this. There is always a positive outcome and the statistic (W) is either less than or equal to one. A value nearer to one suggests a higher likelihood of a regularly distributed set of data. It's crucial to remember that this test is limited to samples that have a size of seven to two thousand.

$$W = \left(\frac{\sum_{i=1}^{n} g_i p_{(i)}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2 \qquad (1)$$

where $p_{(i)}$ is the i_{th} order statistic i.e. the i_{th} smallest number in the sample; $\frac{p_{(1),}p_{(2)},\cdots p_{(n)}}{n}$ mean; the constant ai are given by (Shapiro-Wilk, 1965).

A goodness-of-fit test called the Anderson-Darling test is used to identify if a dataset is representative of a particular probability distribution, like the normal distribution. In its most basic version, it is predicated on the knowledge of the distribution's parameters. This makes the test distribution-free and enables its application without depending on particular values. The Anderson-Darling test, which compares the observed data to a theoretical normal distribution, is more frequently employed to evaluate normality.

$$ADT = -n - \frac{1}{W}\sum_{i=1}^{n}(2_i - 1)\left[\ln P(x_i) + \ln(1 - P(x_i) + \ln(1 - F(x_{n-i+1}))\right] \quad (2)$$

Where
W- size of sample
$P(x)$ is the cumulative distribution function of a specified distribution.

A useful statistical method for determining if data has a normal distribution is the Jarque-Bera test. It evaluates the skewness and kurtosis of the data, which characterize the asymmetry and tail behavior of the distribution. It was created by Carlos Jarque and Anil Bera in 1980. The test can ascertain how well the data fits a typical bell curve by assessing these features.

The test statistics JB is given as

$$JB = \frac{n}{6}\left(S^2 - \frac{(k-3)}{4}\right)^2 \quad (3)$$

Where
N - size of the sample
S - Skewness of the sample
K - kurtosis of the sample

MODEL FOR SIMULATION

Consider an inverse square root response surface model of the form

$$\frac{1}{\sqrt{y_i}} = \beta_0 + \sum_{j=1}^{k}\beta_{jj}x_j + \sum_{j=1}^{k}\beta_j x^2{}_j + \sum_{i<}^{k}\sum_{j=2}^{k}\beta_{ij}x_i x_j + \varepsilon_i \quad (4)$$

Where
$\beta_0, \beta_1 x1, \beta_2 x2, \ldots, \beta_k xk$ are model parameters,
$\frac{1}{y_i}$ is the inverse response variable and
$x_1, x_2, \ldots, x_k$ are factor.

The monte-carlo experiment were conducted as follows; the error terms ε_i were generated. In this study, σ^2 values were obtained and different sample sizes varied between n= 5, 15, 30, 50, 100, 200, 500 and 1000 respectively.

4. DATA ANALYSIS AND RESULTS

4.1 Monte Carlo Simulation Setup

To measure the effect of sample size on quantile based plots for detecting normality, we simulated random numbers following the normal distribution with different mean for various sample sizes (n = 10, 20, 50, 100, 200, 500, 1000) and sets of transformed variables (square root, log, inverse square root and Box-cox). Standard deviations of different sizes were introduced into the simulation in order to test for the effect of dispersion. All analyses were done using R4.1.2

Table 1

n(10)	Transformation	Jarque-berra	shapiro	Anderson
	$\beta_0 = 8$ $\beta_1 = 1.28$ $\beta_2 = 1.34$			
	log	0.795	0.6997	0.7411
	square root	0.762	0.603	0.6236
	box-cox	0.6988	0.4457	0.4741
	inverse-square root	0.0765	0.0823	0.45871
n(20)	**Transformation**	**Jarque-berra**	**shapiro**	**Anderson**
	$\beta_0 = 10$ $\beta_1 = ,0.86$ $\beta_2 = 0.45$			
	log	0.745	0.6297	0.6088
	square root	0.599	0.3862	0.3935
	box-cox	7.93E-06	3.96E-09	7.49E-14
	inverse-square root	0.8195	0.97049	0.7532

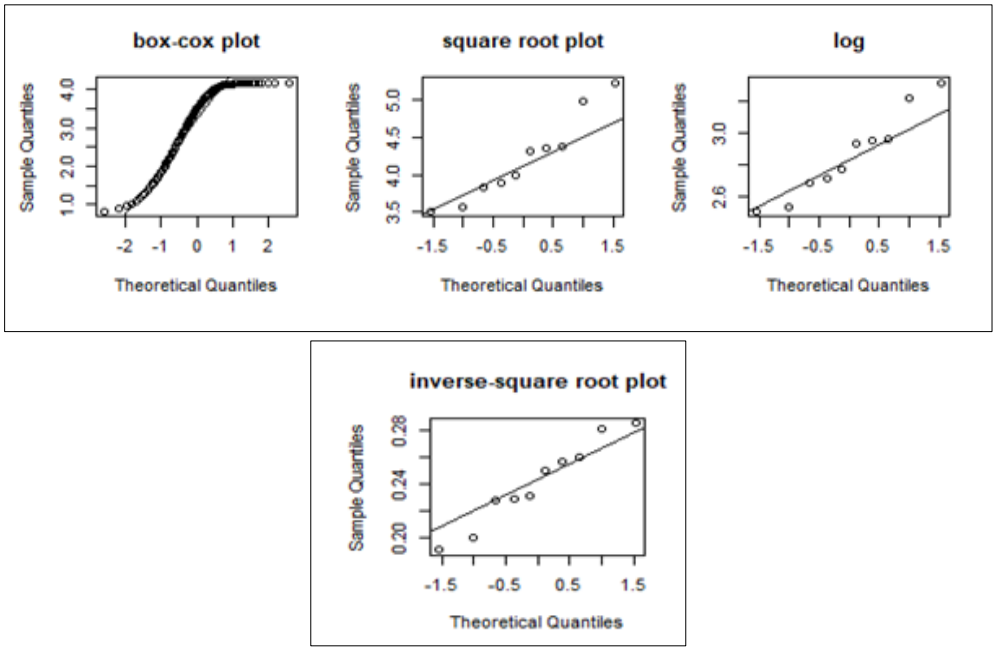

Fig 1: Quantile plots of transformed data when n=10

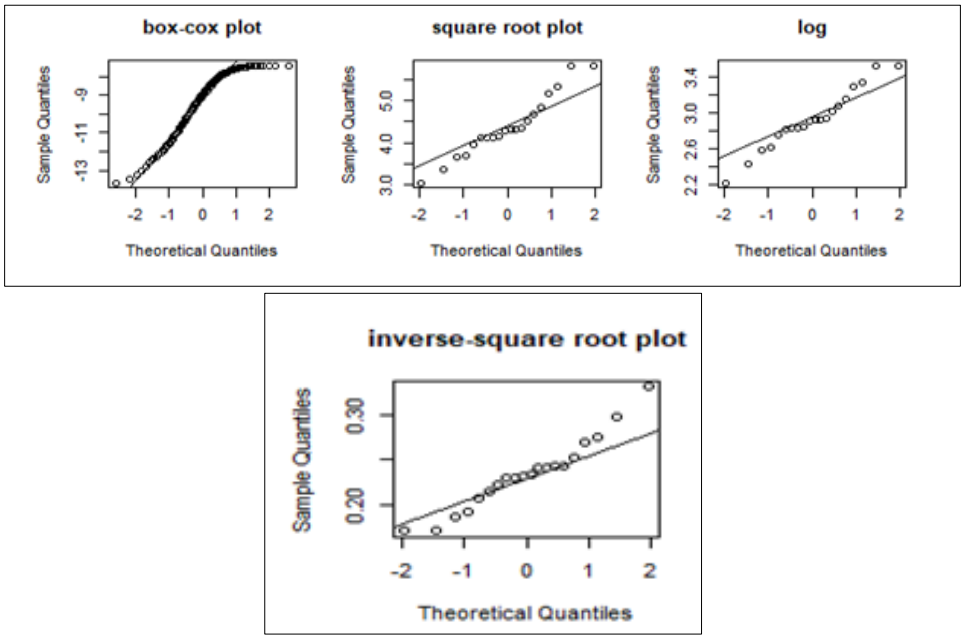

Fig 2: Quantile plots of transformed data when n=20

Table 2: Normality tests on transformed data when n = 50 and n = 100

n(50)	Transformation	Jarque-Berra	shapiro	Anderson
	$\beta_0 = 15$ $\beta_1 = 2.86$ $\beta_2 = 3.45$			
	log	0.3992	0.1548	0.03849
	square root	0.4947	0.2515	0.06199
	box-cox	0.001613	1.68E-08	1.02E-12
	inverse-square root	0.2969	0.08106	0.02205
n(100)	**Transformation**	**Jarque-Berra**	**shapiro**	**Anderson**
	$\beta_0 = 25$ $\beta_1 = 6.86$ $\beta_2 = 8.45$			
	log	0.1471	0.1476	0.1788
	square root	0.4963	0.7321	0.7052
	box-cox	0.01049	2.69E-06	1.67E-07
	inverse-square root	0.08966	0.066933	0.0153

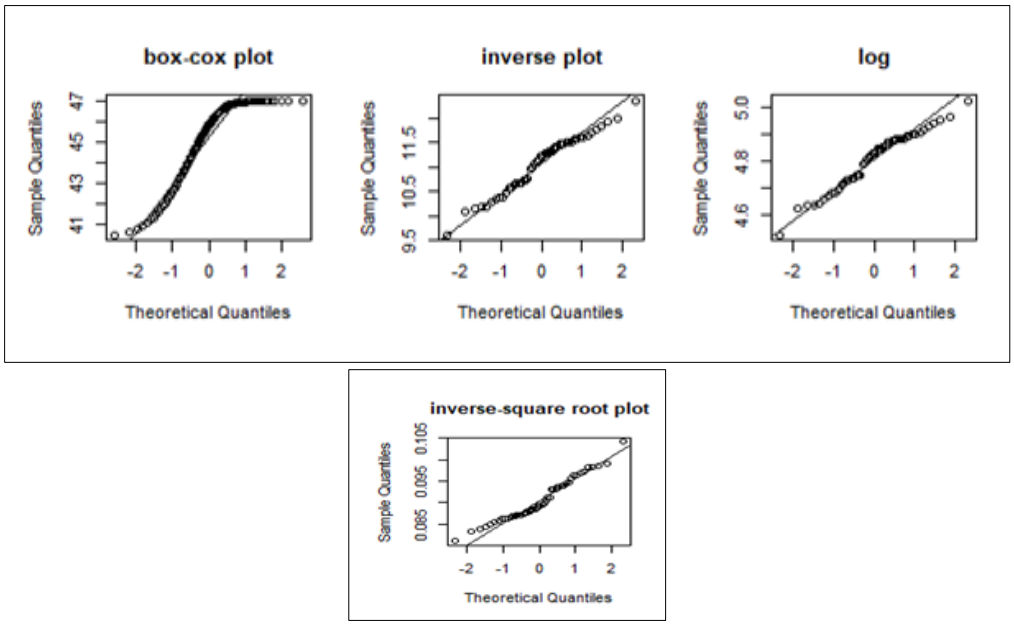

Fig 3: Quantile plots of transformed data when n=50

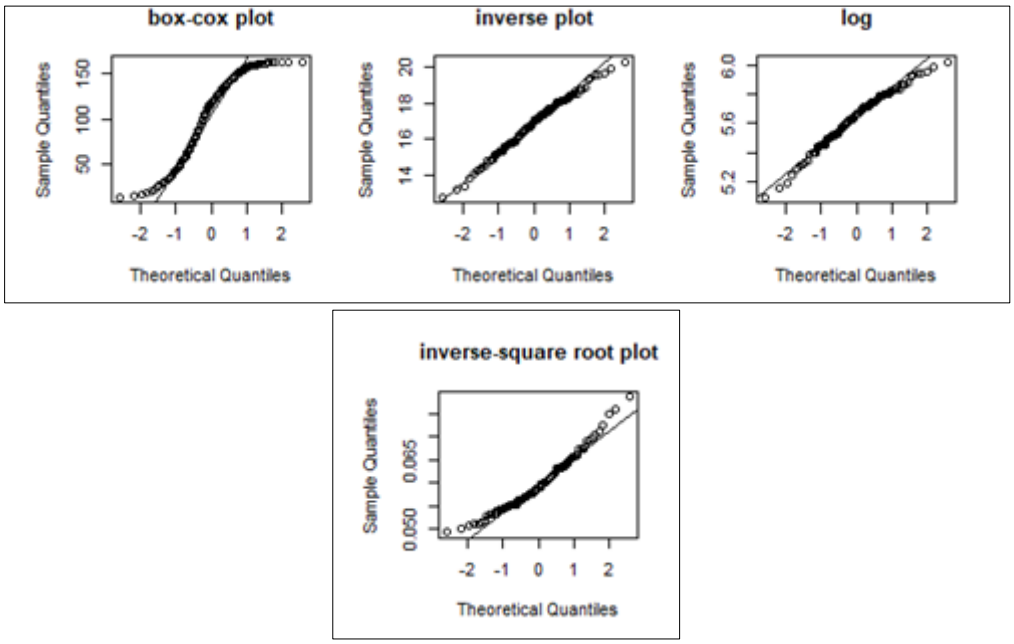

Fig 4

Table 3: Normality tests on transformed data when n = 200 and n = 500

n(200)	Transformation	Jarque-Berra	shapiro	Anderson
	$\beta_0 = 25$ $\beta_1 = 6.86$ $\beta_2 = 8.45$			
	log	0.7195	0.4777	0.1152
	square root	0.902	0.5918	0.304
	box-cox	0.009492	1.60E-06	5.28E-08
	inverse-square root	0.1148	0.06449	0.01282
n(500)	**Transformation**	**Jarque-Berra**	**shapiro**	**Anderson**
	$\beta_0 = 45$ $\beta_1 = 14.68$ $\beta_2 = 11.45$			
	log	4.67E-12	2.04E-06	4.11E-05
	square root	0.1351	0.1111	0.02656
	box-cox	0.01267	5.71E-06	7.15E-07
	inverse-square root	0.220	0.563E	0.111

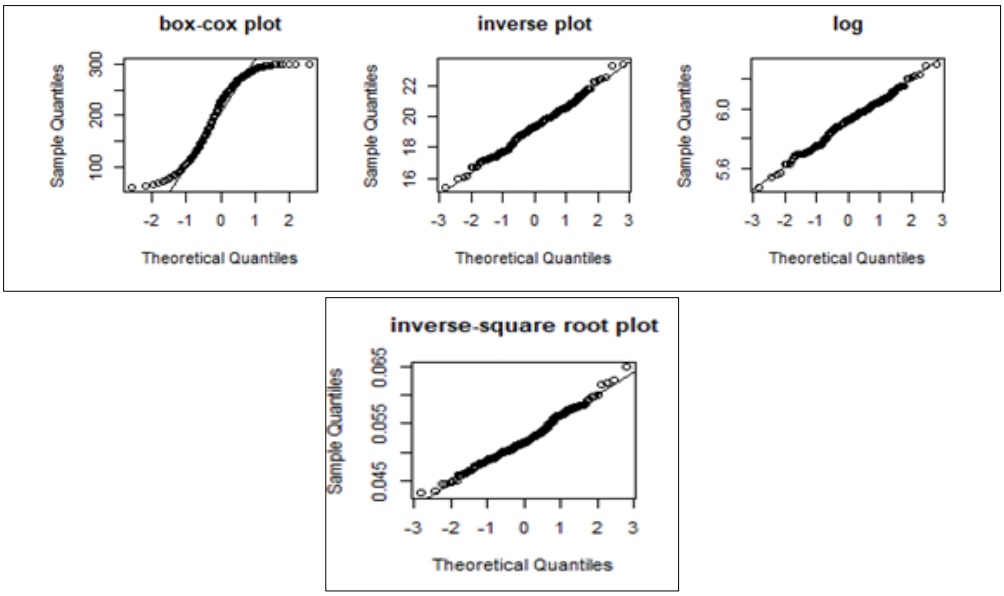

Fig 5: Quantile plots of transformed data when n=200

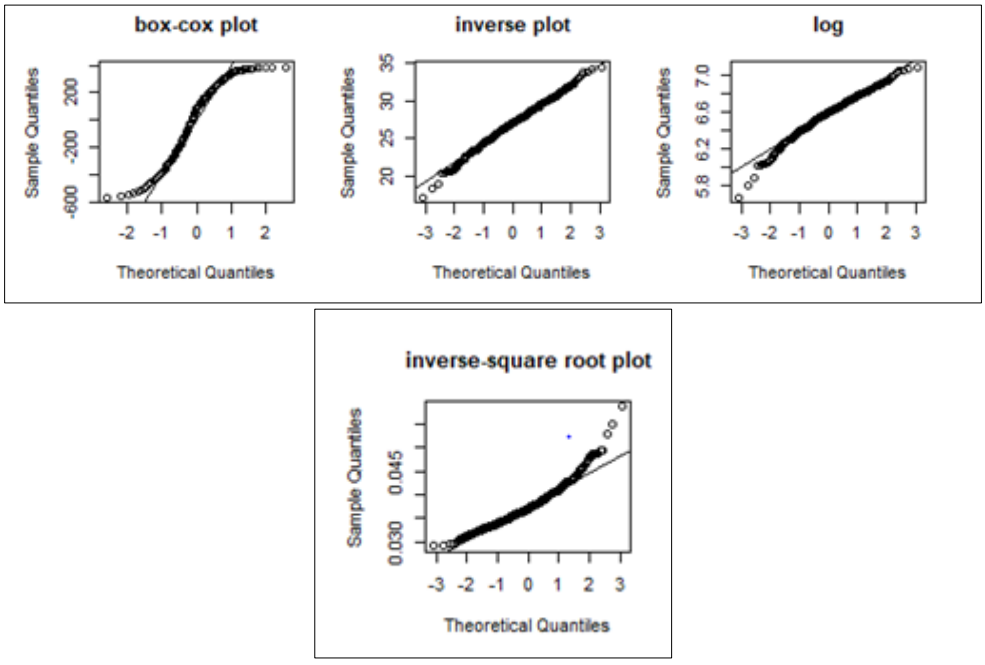

Fig 6: Quantile plots of transformed data when n=500

Table 4: Normality tests on transformed data when n = 1000

n(1000)	Transformation	Jarque-berra	shapiro	Anderson
	$\beta_0 = 4.5$ $\beta_1 = 1.443$ $\beta_2 = 2.45$			
	log	0.03734	0.01764	0.111
	square root	5.73E-06	0.0003922	0.0007938
	box-cox	0.01378	8.13E-06	1.36E-06
	inverse-square root	0.05068	0.1217	0.1954

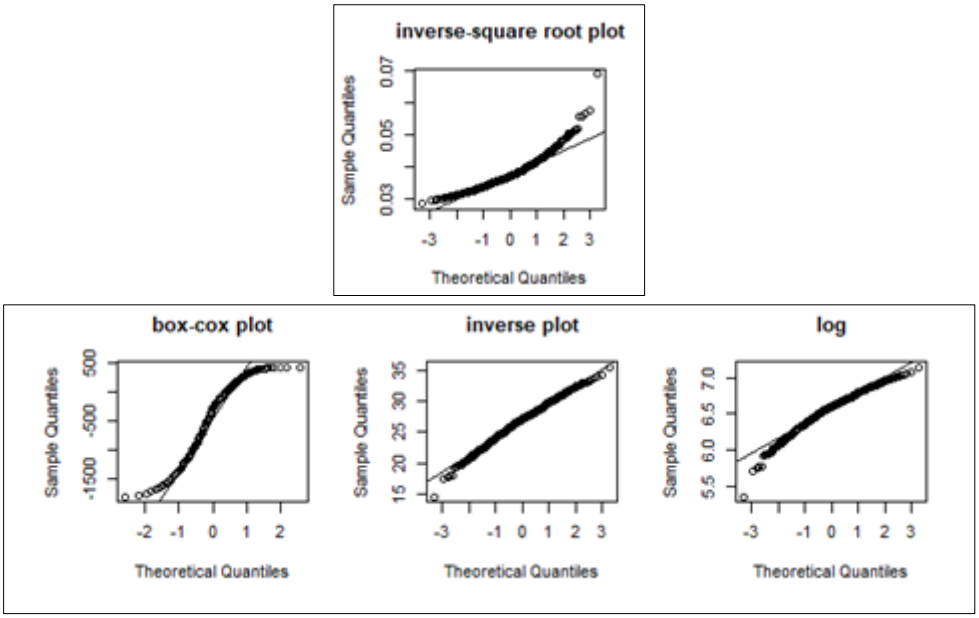

Fig 7: Quantile plots of transformed data when n=1000

5.0 DISCUSSION OF RESULTS

It is discovered that normality increases as the sample sizes increase on the quantile plots. The transformed data is used to compare the normality with the introduction of Anderson-Darlington, Shapiro-wilk and Jarque-Berra test statistics. It is also discovered that dispersion has effect on normality of the transformed data, the departure from normality increases as the measure of dispersion increases.

Researchers have compared different normality tests and found that square root and inverse square root transformations are surprisingly strong contenders. While the Shapiro-Wilk test, with its adjustment for covariance, might seem theoretically superior, both approaches perform about equally well in practice. Interestingly, the square root and inverse square root transformations might even be slightly more powerful in detecting certain deviations from normality.

6.0 CONCLUSION AND RECOMMENDATION

Drawing valid conclusions about whether data is normally distributed is a critical task for researchers worldwide, often achieved through plots and classical tests. In this study, the plot of squared transformed data for n=1000 might suggest normal distribution visually, with square root and inverse square root transformations showing significance across sample sizes, but box-cox indicating rejection in some cases. Quantile plots appear most suitable for smaller samples.

The results indicate that sample sizes and measures of dispersion significantly impact the detection of normality using quantile plots. Normality tends to increase with larger sample sizes but decreases with higher standard deviations. Therefore, it is advisable to use quantile plots and inverse square root transformations for small sample sizes, while classical tests offer objective and precise results across all sample sizes.

REFERENCES

- Aihounton, G. B., & Henningsen, A. (2021). Units of measurement and the inverse hyperbolic sine transformation. *The Econometrics Journal*, *24*(2), 334-351.
- Cromwell, J. B., Labys, W. C., & Terraza, M. (1994). *Univariate Tests for Time Series Models*, Sage, Thousand Oaks, CA, 20–22.
- Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*(6), 2576-2590.

- Lee, D. K. (2020). Data transformation: a focus on the interpretation. *Korean journal of anesthesiology*, *73*(6), 503.
- Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annu Rev Public Health*, *23*, 151–169.
- Peterson, R. A. (2021). Finding Optimal Normalizing Transformations via best Normalize. *R Journal*, *13*(1).
- Royston, J. P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *31*(2), 115-124.
- Shapiro, S. S., & Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American statistical Association*, *67*(337), 215-216.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591-611.