International Journal of Advanced Statistics and Probability, 12 (1) (2025) 17-27

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
https://doi.org/10.14419/8bnsc148
Research paper

Diagnosing and correcting violations of normality and constant variance assumptions in multiple linear regression analysis

Victor Chijindu Iheaka *

Department of Statistics, Imo State University, Owerri, Imo State, Nigeria
**Corresponding author E-mail: victorgodwill@ymail.com*

Received: February 19, 2025, Accepted: March 6, 2025, Published: March 20, 2025

Abstract

This study showcased the significance of correcting for Non-normality and Nonconstant variance of residuals in linear regression modelling. The concept was demonstrated by using two different hypothetical, Data M (the Initial Dataset) and Data N (the Initial Dataset). The diagnosis of non-normality and nonconstant variance was performed using the Anderson-Darling test (or D'Agostino Omnibus test) and White test, respectively, revealing their presence in the models for the initial datasets, while the assumptions of no multicollinearity and no autocorrelation were met. The model established for Data M (the Initial Dataset) was statistically significant with an R-square value of 0.538, an AIC value of 1071.424, an SBC value of 1083.787, and an RMSE value of 9774.849. Similarly, the model established for Data N (the Initial Dataset) was statistically significant with an R-square value of 0.865, an AIC value of 768.443, an SBC value of 776.427, an RMSE value of 584.946. Data M (the Initial Dataset) and Data N (the Initial Dataset) were transformed using the Semi-Logarithm transformation method, generating new sets of data, Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected). After the correction was made, the datasets complied with all the linear assumptions necessary for regression analysis. The multiple linear regression model estimated for Data M (Non-normality and Nonconstant Variance Corrected) was found to be statistically significant, achieving an R-square value of 0.748, an AIC value of -81.061, an SBC value of -61.699, and an RMSE value of 0.473; and the model established for Data N (Non-normality and Nonconstant Variance Corrected) was statistically significant with an R-square value of 0.871, an AIC value of -145.907, an SBC value of -137.529, an RMSE value of 0.287. Based on the R-squares, AIC, SBC, and RMSE values for both initial and transformed models, it was concluded that the estimated regression model for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) demonstrated superior model performance when compared to the regression models for Data M (the Initial Dataset) and Data N (the Initial Dataset).

Keywords: *Linear Regression Analysis; Multiple Linear Regression Model; Residuals; Non-Normality Assumption; Nonconstant Variance Assumption; Correcting.*

1. Introduction

Conducting normality and constant variance assumptions are very essential not only in a linear regression analysis but other statistical analysis that deals with parametric tests. Most parametric tests are called for meeting the assumptions of normality and constant variance. Kim and Park (2009) explained that when the data satisfies the normality assumption, it shows a probability distribution curve with highest frequency of occurrence at the center, and the frequency decreases with distance from the center which can be regarded as a bell-shaped distribution with a zero mean and standard deviation of one (1). According to Barker and Shaw (2015), for the regression estimates obtained from Ordinary Least Square (OLS) method to produce valid confidence intervals and P-values, the residuals must be independent, be normally distributed and have a constant variance. If these assumptions are not satisfied, the regression estimates will remain unbiased but no longer be the Best Linear Unbiased Estimator (BLUE) and the forecasting power will be reduced.

Notably, testing for normality is a crucial assumption in linear regression analysis and it explicitly involves checking the normality of the residuals (the differences between observed and predicted values), not the original data itself. Carrying out statistical tests for independence and constant variance of the residuals, and correcting their existence in the regression model, are crucial. However, the existence or non-existence of multicollinearity in the data might be overlooked if the primary purpose of the linear regression model parameters is forecasting, as highlighted in the literature (see, for example, Koutsoyiannis, 1977). In a situation where the residuals from a linear regression analysis do not follow a normal distribution, it is referred to as non-normality while the variance of residual is not the same for all observations is regarded to be nonconstant variance.

Generally, the presence of outliers in the data, skewed data, sampling issues, incorrect data transformations, incorrect functional form, model misspecification and measurement errors are the causes of non-normality and nonconstant variance of residuals in the dataset. Gujarati (2004) explained that a data point that is significantly different from other data points in a dataset, regression model where some important variables are omitted from the model, ratio or first difference transformation and incorrect functional form are the sources of nonconstant variance of residuals existing in the Linear Regression Model (LRM). Gujarati (2004) further explained that the problem of nonconstant variance of residuals is likely to be more common in cross-sectional data than in time series data.

Interestingly, the logarithmic transformation tends to be popular along with other "variance stabilizing" transformation such as the square root and power transformations of data targeted to correct the existence of nonconstant variance of residuals in the model as have been shown in some of the literatures (see, for examples, Osaro, 2018; Ohaegbulem and Iheaka, 2024). Jude and Isobeye (2021) suggested that the nonparametric Theil's simple linear regression is an alternative to OLS when there is existence of non-normality in a data, the method should be employed to remedied the situation. Jude and Isobeye (2021) further suggested that if this assumption still fails to hold after attempting various remediation methods, one should check for outliers. If outliers are detected, they should be removed, and the underlying assumption should be re-examined again. Other methods of remedying the nonconstant variance of residuals in the LRM includes; robust estimation methods for standard errors (see, for example, White, 1980), bootstrap methods (see, for example, Flachaire, 2005) and Weighted Least Squares method (see, for example, Pedace, 2013).

According to Hogg (1979), regression estimates can be especially sensitive to heavily tailed distribution. The Gaussian-Markov theorem demonstrated that in linear regression analysis, the estimates obtained from OLS are linear unbiased estimates with smallest possible variance even if the residuals are not normally distributed (that is, to said that normality assumptions is robust to large dataset). This theorem may sometimes be misinterpreted to showcased that normality assumptions might not be important when carrying out a linear regression analysis on a large sample data (asymptotically). The Gaussian-Markov theorem only concerns point estimates not the confidence intervals or t-test. Thus, even if the data is large, the normality assumption are called for to be tested as some literatures have shown the need for testing normality than disregarding the assumption when the sample data is large (see, for examples, Judge et al., 1985; Koenker, 1982; Das and Imon, 2016). In addition, Hawkins (1989) stated that violations of normality might be especially problematic for inferences based on correlations, for which even large sample sizes are unlikely to help.

Consequently, when a Linear Regression Model (LRM) fails to meet the assumptions of normality and constant variance of residuals, despite meeting other assumptions, the usual tests of statistical significance, such as t and F tests, become invalid. This leads to inflated Type 1 error rates, and the regression estimates are no longer the best linear unbiased estimates. Also, the standard errors of such regression estimates are biased and inconsistent, which may result in unreliable future forecasting of the dependent variable. Therefore, it is crucial to test for and correct non-normality and nonconstant variance of the residuals when identified. This ensures the reliability of the estimated regression coefficients, forecasting power, and validity of statistical inference.

This paper aims to demonstrate methods for testing the normality and constant variance of residuals in Multiple Linear Regression Model (MLRM), identifying any violations and presenting a method for their correction. Specifically, the objectives of this study are as follows; to demonstrate the methods of testing for non-normality and nonconstant variance of residuals in a MLRM, to showcase the existing method of correcting it and to determine with a view to comparing the distinctive of the MLRM that consist of non-normality and nonconstant variance of residuals and when it has been corrected.

2. Literature review

Some studies have been carried out in the past which centered on various methods of correcting non-normality, and also different methods of correcting nonconstant variance of residuals in LRM. Here are reviews of some of these studies:

Osemeke et al. (2024) studied detection and correction of violations of linear model assumptions by means of residuals. The data from a bread bakery in Nigeria whose interest was to established the relationship between the effect of trademark (X_1), bread texture (X_2) and bread aroma (X_3) on Consumers' attitude, Y were analysed using multiple regression analysis. White test was used to verify the existence on nonconstant variance of residuals in the model. The results showed that the P-value for this White test was 0.043 which implied that nonconstant variance of residuals were present in the model. Anova F-test was used to assessed the non-normality, the results obtained demonstrated that the P-values for X_1, X_2 and X_3 were 0.855, 0.892 and 0.083, respectively, indicated that the individual variables showed evidence of non-normality. Means of residuals obtained from the model was used for transformations, and the P-values for White test after transformations was 0.381 which means that the model has been corrected for nonconstant variance of residuals. In the same vein, Anova results for assessing non-normality implied that non-normality was corrected after transformations. It was concluded that this method of means of residuals for correcting the violations of LRM assumptions showcased ability to correct it where it existed in the model.

Thinh et al. (2020) carried out research on linear regression models for heteroscedastic and non-normal data. A simulation dataset with sample sizes of 20, 50 and 100 were analyzed using several methods proposed to handle problems of heteroscedastic and non-normal of residuals, so that proper investigations of the performance of these several methods of estimation used in this study will be reviewed. These several methods of estimation included; ordinary least squares (OLS), Transform both Sides (TBS) regarded as logarithm transformation, Power of the Mean Function (POM) and Exponential Variance Function (VEXP) which was used to handle the three different kinds of the nonconstant variances under four symmetric distributions. Relative bias, Mean Squared Error (MSE) and coverage probability of the nominal 95% confidence interval for regression parameters were all assessed. The simulation results and application to real life data suggest that each estimation method performed differently on different variance structures and different distributions whereas the sample size did not give much effect on each estimation method. It was concluded that the TBS method (log-transformation) performed best in terms of smallest bias and MSE, especially under extreme heteroscedasticity compared to POM and VEXP whereas the OLS method obviously overestimates the slope parameter.

Ohaegbulem and Iheaka (2024) conducted a study involving the remedying the presence of heteroscedasticity in a multiple linear regression modelling. The two hypothetical datasets were analysed using multiple regression and correlation analyses. The results for testing heteroscedasticity for Data A (the Original) and Data B (the Original) before the log-transformation using BPG test confirmed statistically significant at 5% level of significance (P-values = 0.0025 and 0.0091, respectively), which implied there is existence of heteroscedasticity in the both original models. Log transformation of the data was employed in order to remedied the existence of heteroscedasticity in the original models. The result for testing heteroscedasticity for Data A (Now with Heteroscedasticity Remedied) and B (Now with Heteroscedasticity Remedied) after the log-transformation using BPG test confirmed statistically insignificant at 5% level of significance (P-values= 0.3134 and 0.1226, respectively), which implied there is existence of homoscedasticity in the both transformed models. The values of the R^2 for Data A (Now with Heteroscedasticity Remedied) and Data B (Now with Heteroscedasticity Remedied) are 0.986 and 0.624,

respectively, greater than the values of the R^2 for Data A (the Original) and Data B (the Original) which are 0.976 and 0.553, respectively. Also, the values of the AIC for Data A (Now with Heteroscedasticity Remedied) and Data B (Now with Heteroscedasticity Remedied) are -135.02 and -120.36, respectively, lesser than the values of the AIC for Data A (the Original) and Data B (the Original) which are 332.59 and 347.25, respectively. It was concluded Log-transformation of the variables yielded better models estimates and statistics than the regression models for Data A (the Original) and B (the Original).

3. Materials and methods

For illustrative purposes, this research utilized two hypothetical datasets, designated as Data M and Data N, to contextualize the study. Data M (the Initial Dataset) consists of five (5) independent variables and one dependent variable, summarized in Columns 1 to 6 of Table 3.1 (see Appendix A); while Data N (the Initial Dataset) consists of three (3) independent variables and one dependent variable, summarized in Columns 1 to 4 of Table 3.2 (see Appendix B).

The Multiple Linear Regression (MLR) analysis is used to establish the relationship that exists among a dependent variable and a set of related independent variables. Through the utilization of the OLS procedure, a statistical analysis technique, the coefficients for the independent variables are estimated. Specifically, in this study, after establishing the relationship between the dependent and independent variables, model evaluation metrics will be adopted to assess the predictive accuracy of the dependent variable. The multiple linear regression model, a statistical framework that elucidates the relationship among the dependent and independent variables, is commonly represented by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u_i \tag{1}$$

Equation (1) can also be expressed in matrix terms (see, for example Kurtner et al. 2005) as,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{pmatrix} \tag{2}$$

For ease of exposition, one can write (2) as,

$$\underset{(n\times 1)}{Y} = \underset{(n\times k)}{X}\underset{(k\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon} \tag{3}$$

Applying the OLS method (see, for example Kurtner et al. 2005) the regression model parameters, are $\hat{\beta}_i$'s estimated as,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{4}$$

Then, the estimated regression model will be obtained by substituting the values of the $\hat{\beta}_i$'s in (4) into (1).

After establishing the multiple linear regression model, it is generally recommended to test the assumptions of linear regression analysis first, before evaluating its predictive accuracy using the coefficient of multiple determination, R^2, the Akaike Information Criterion (AIC), the Schwarz Bayesian Criterion (SBC) and Root Mean Square Error (RMSE).

Furthermore, though, Normality, Constant variance, Autocorrelation and Multicollinearity are the most prominent assumptions that are supposedly to be met, so that the estimated regression model parameters will be reliable. Nevertheless, it does not imply that the other assumptions of linear regression analysis are of less importance.

3.1. Tests for the assumptions of linear regression analysis

Traditionally, testing the assumptions of linear regression analysis is considered a preliminary step, preceding the actual regression analysis. This informal assessment enables a timely evaluation of how well the model fits the data. However, a counterintuitive conclusion arises: some tests for normality, autocorrelation and constant variance assumptions require the residuals from the established model before they can be tested. Consequently, it is necessary to conduct the regression analysis first to obtain the residuals, which are essential for evaluating these assumptions. Therefore, a pragmatic approach is to perform the regression analysis, obtain the residuals, and then test the assumptions, addressing any issues that arise to ensure the reliability of the regression estimates are obtained from the model.

a) Test for the normality assumption

One of the assumptions of linear regression analysis required by the OLS method for the estimability of the parameters in the regression model is that the residuals are normally distributed. Meanwhile, in order to confirm the existence of normality in a regression model, two methods are adopted, namely; informal method and formal method are explained as follows:

Informal Method: Gujarati (2004) stated that a simple graphical representation can be used to explain whether the residuals are normally distributed. So, two common graphical methods for assessing normality of residuals are the histogram of residuals and the normal probability plot (also known as the P-P plot or Q-Q plot). The histogram of residuals is simply computed by plotting the values of expected residuals against the random variable which will produce erect rectangles equal in height to the number of observations and the shape of normal distribution curve can be ascertained on the histogram.

Formal Method: Apart from the graphical inspection of normality test, in order to confirm the existence of normality in the dataset, some formal methods of testing the normality assumption can be used which includes; Shapiro-Wilk test, Anderson-Darling test, Kolmogorov test and D'Agostino Omnibus test. The Anderson-Darling test will be used to study the shape of the probability density function of the random variables. D'Agostino et al. (1990) describes a normality test that combines the tests for skewness and kurtosis, for example, the null hypothesis of a normally distributed residuals is to be rejected if and only if the D'Agostino Omnibus, K^2, is greater than or equal to

the critical value, $\chi^2_{\alpha;2}$ (that is, H_0 is to be rejected if and only if $K^2 \geq \chi^2_{critical}$). The D'Agostino Omnibus test statistic (see, for example, D'Agostino et al., 1990), is given by,

$$K^2 = Z_S^2 + Z_k^2 \tag{5}$$

Where,

$$Z_s = \delta \ln\left(\frac{Y}{\alpha} + \sqrt{\frac{Y^2}{\alpha^2}+1}\right) \tag{6}$$

And

$$Z_k = \frac{\left(1-\frac{2}{9A}\right) - \left(\frac{1-\frac{2}{A}}{1+G\sqrt{\frac{2}{A-4}}}\right)^{1/2}}{\sqrt{\frac{2}{9A}}} \tag{7}$$

The critical value for the D'Agostino Omnibus normality test is given by $\chi^2_{\alpha;2}$; where, 2 is the degree of freedom. The critical value can be read off from a statistical table, such as Neave (1978).

b) Test for homoscedasticity assumption

Constant variance of the residuals is a critical assumption of the ordinary least square method, ensuring the reliable estimation of regression model parameters. According to Gujarati (2004), in order to confirm the existence of heteroscedasticity, two methods are used, namely; Informal method and Formal method are explained as follows:

Informal Method: This method uses graphical approach for detection of the presence of heteroscedasticity in a linear regression model. Gujarati (2004) explained that the estimated squared residual, $\hat{u}_i^2$ are plotted against the predicted values, $\hat{Y}_i$ or the each of the independent variables. Gujarati (2004) further explained that the idea of this graphical presentation is to find out whether the estimated mean value of observed values, Y is systematically related to $\hat{u}_i^2$ or any systematic pattern between any regressor and the residuals confirms the insignificant test of homoscedasticity.

Formal Method: Applying formal methods of testing constant variance of residuals, some commonly used tests are namely; Breusch-Pagan test, Spearman Rank Correlation test, Goldfeld-Quandt test, Park test, Glejser test, White test and Koenker-Basset test. In the literature, the most and frequently used method of testing constant variance assumption is White test due to its does not depends on omission of the values, graphical inspection of the residuals and not sensitive to normality assumption. According to Gujarati (2004), for example, the null hypothesis which states that the error terms are homoscedastic is to be rejected if and only if the calculated White test statistic, LM, is greater than or equal to the critical value (that is, H_0 is to be rejected if and only if LM $\geq \chi^2$ -critical). The White test statistic (see, for example, White, 1990), is given by,

$$LM = n.R^2 \tag{8}$$

Where,

n is the total number of observations, and R^2 is obtain from the auxiliary regression (that is, the squared residuals from the original regression is regressed on the original independent variables, their squared values and the cross product(s) of the independent variables). The critical value for the White test is given by $\chi^2_{\alpha;k}$; where, k is the degree of freedom. This critical value can be read off from a statistical table, such as Neave (1978).

3.2. Mitigation of violated assumptions in linear regression analysis

In forecasting models, two crucial assumptions of linear regression analysis that require remediation when violated are independence and constant variance of residuals. Their existence in a linear regression model does not destroy the unbiasedness and consistency properties of the OLS estimators, but they are no longer efficient, not even asymptotically. Gujarati (2004) stated that this lack of efficiency makes the outcome of the usual hypothesis-testing to be dubious. Kurtner et al. (2005) explained that the data transformations will be helpful in eliminating the problem of dependence and nonconstant variance of the residuals. One should adopt appropriate remedial actions, such as transformation of the variables, deletion of outliers, etc., to address the problems caused by violations of these assumptions. Bartlett (1947) postulated that transforming variables (for examples, arcsine, power and logarithm transformations) can simultaneously address the problem of non-normality and nonconstant variance, promoting the reliability and accuracy from the model.

However, to validate Bartlett's (1947) claim, this study will employ a semi-logarithmic transformation method to simultaneously address non-normality and nonconstant variance of residuals. The procedure of this semi-logarithmic transformation method involves taking the square root of each entry in the dataset and then applying the natural logarithm (see, for example Nwankwo 2011). After the data transformation one can now apply the OLS method to estimate the coefficients in the equation. The semi-logarithm transformation for multiple linear regression analysis is given as,

$$In\sqrt{Y_i} = \beta_0 + \beta_1 In\sqrt{X_{1i}} + \beta_2 In\sqrt{X_{2i}} + \cdots + \beta_k In\sqrt{X_{ki}} + u_i \tag{9}$$

Also, this semi-logarithm transformation can be referred as log-transformation with square roots. Generally, semi-logarithm transformation is often used to stabilize variance and normalize data.

4. Results and discussion

Adopting (1), this study uses the following theoretical model to assess the independent variables that are associated with the dependent variable; for Data M (the Initial Dataset) and Data N (the Initial Dataset), the multiple linear regression equations will, respectively, be given as,

$$Y_M = \beta_{0M} + \beta_{1M} X_{1M} + \beta_{2M} X_{2M} + \beta_{3M} X_{3M} + \beta_{4M} X_{4M} + \beta_{5M} X_{5M} + u_M \quad (10)$$

And

$$Y_N = \beta_{0N} + \beta_{1N} X_{1N} + \beta_{2N} X_{2N} + \beta_{3N} X_{3N} + u_N \quad (11)$$

The data analyses in this study shall be done with the aid of the following statistical packages; Microsoft Office Excel (2021), Minitab (2019), SPSS version 26, and NCSS (2024). The results outputs from the various computer packages employed in testing the relevant assumptions of the multiple linear regression analysis, as well as the main data analyses are summarized in Tables 4.1 to 4.14 and Figs. 4.1 to 4.4.

The procedure of carrying out the multiple linear regression analysis, starting from the tests of assumptions to the establishment of the multiple linear regression model for Data M (the Initial Dataset) and Data N (the Initial Dataset) are as presented in Table 4.1 to 4.7 and figure 4.1 to 4.2.

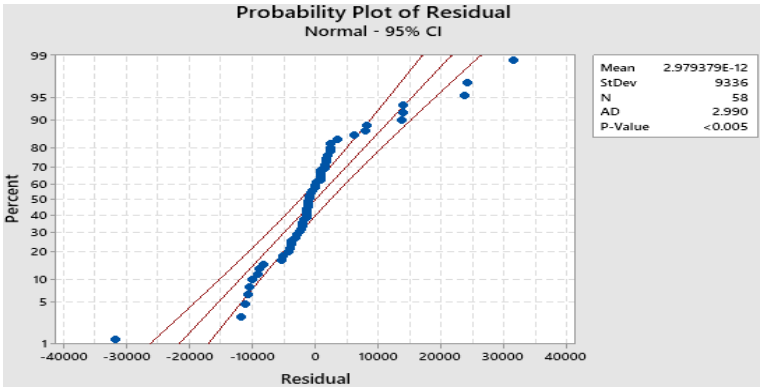

Fig. 4.1: The Anderson-Darling Test for the Normality Assumption on Data M (the Initial Dataset).

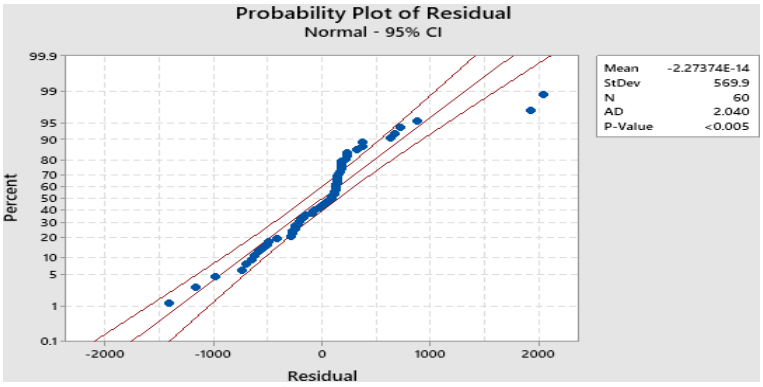

Fig. 4.2: The Anderson-Darling Test for the Normality Assumption on Data N (the Initial Dataset).

Table 4.1: D'Agostino Omnibus Test for Normality Assumption on Data M (the Initial Dataset and Data N (the Initial Dataset)

	Data M (the Initial Dataset)	Data N (the Initial Dataset)
K^2-stat	15.0584	19.7504
p-value	0.0005	0.0001
alpha	0.05	0.05
normal	no	no

Table 4.2: White Test for Constant Variance Assumption on Data M (the Initial Dataset) and Data N (the Initial Dataset)

	Data M (the Initial Dataset)	Data N (the Initial Dataset)
LM stat	20.236	20.049
df	2	2
p-value	4.03E-05	4.43E-05
F stat	14.736	14.303
df1	2	2
df2	55	57
p-value	7.5E-06	9.25E-05

Table 4.3: Regression Model Coefficients for Data M (the Initial Dataset)

Variable	Unstandardized Coefficients	Standardized Coefficients	t-stat	Sig.	Correlations	Collinearity Statis-tics

	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
β_{0M}	-8005.967	5129.225		-1.561	0.125					
X_{1M}	0.001	0.000	0.585	5.455	0.000	0.682	0.603	0.514	0.771	1.297
X_{2M}	-1.991	2.048	-0.129	-0.972	0.336	-0.061	-0.134	-0.092	0.501	1.997
X_{3M}	1.672	2.267	0.099	0.738	0.464	-0.075	0.102	0.070	0.496	2.016
X_{4M}	57.618	93.106	0.066	0.619	0.539	0.317	0.086	0.058	0.788	1.269
X_{5M}	23485.949	11452.816	0.231	2.051	0.045	0.508	0.274	0.193	0.698	1.434

Table 4.4: Regression Model Coefficients for Data N (the Initial Dataset)

Variable	Unstandardized Coefficients		Standardized Coefficients	t-stat	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
β_{0N}	-708.366	474.785		-1.492	0.141					
X_{1N}	1.378	0.243	0.491	5.666	0.000	0.889	0.604	0.278	0.321	3.115
X_{2N}	0.085	0.016	0.457	5.266	0.000	0.883	0.576	0.258	0.320	3.129
X_{3N}	0.614	0.486	0.066	1.263	0.212	0.382	0.166	0.062	0.878	1.139

Table 4.5: Model Summary for Data M (the Initial Dataset)

Multiple R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
				R Square Change	F Change	df1	df2	Sig. F Change	
0.734	0.538	0.494	9774.8485126	0.538	12.119	5	52	0.000	1.814

Table 4.6: Model Summary for Data N (the Initial Dataset)

Multiple R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
				R Square Change	F Change	df1	df2	Sig. F Change	
0.930	0.865	0.858	584.94614	0.865	119.701	3	56	0.000	2.030

Table 4.7: Additional Overall Fit of the Regression Models for Data M (the Initial Dataset and Data N (the Initial Dataset)

	Data M (the Initial Dataset)	Data N (the Initial Dataset)
AIC	1071.424	768.443
SBC	1083.787	776.820
RMSE	9774.849	584.946
Ave. Abs. PCT Error	450.376	163.427

From the normal probability plot and the Anderson-Darling (AD) test in Figs. 1 and 2, which were used to verify that the residuals are normally distributed, the computed Anderson-Darling statistic, AD, for Data M (the Initial Dataset) and Data N (the Initial Dataset) are 2.990 and 2.040, respectively, with (P=0.005 and 0.005, respectively); which are less than the level of significance, $\alpha = 0.05$. Therefore, the null hypothesis (which states that residuals are normally distributed) is rejected; thus, the conclusion is that the residuals are not normally distributed. Also giving support to this conclusion are the D'Agostino-Omnibus test (in Table 4.1) for Data M (the Initial Dataset) and Data N (the Initial Dataset) which are 15.058 and 19.750, respectively, with (P=0.0005 and 0.0001, respectively); which are less than the level of significance, $\alpha = 0.05$.

White test in Table 4.2 was used to test for the constant variance assumption, the value of the computed test statistic for Data M (the Initial Dataset) and Data N (the Initial Dataset) are 20.236 and 20.049, respectively, with (P=0.000 and 0.000, respectively); which are less than the level of significance, $\alpha = 0.05$. Therefore, the null hypothesis (which states that the residuals are homoscedastic) is rejected; thus, the conclusion is that the residuals are heteroscedastic.

Despite Data M (the Initial Dataset) and Data N (the Initial Dataset) failing the constant variance assumption, the multiple linear regression analysis was still carried out on the both hypothetical datasets. From Tables 4.3 and 4.4, the multiple linear regression models for Data M (the Initial Dataset) and Data N (the Initial Dataset), respectively, are obtained as,

$$\hat{Y}_M = -8005.97 + 0.001X_{1M} - 1.991X_{2M} + 1.672X_{3M} + 57.618X_{4M} + 23485.949X_{5M} \quad (12)$$

And

$$\hat{Y}_N = -708.366 + 1.378X_{1N} + 0.085X_{2N} + 0.614X_{3N} \quad (13)$$

Also shown in Table 4.3, the VIF values < 10 for all five independent variables in Data M (the Initial Dataset), indicating no multicollinearity issues; and Table 4.4 shows VIF values < 10 for all three independent variables in Data N (the Initial Dataset), indicating no multicollinearity issues. From Tables 4.5 and 4.6, the value of the computed Durbin-Watson statistic for Data M (the Initial Dataset) and Data N (the Initial Dataset) are 1.8 and 2.0, respectively, (which are approximately equal to 2) implies that there is no existence of autocorrelation in the models. Furthermore, Tables 4.5 and 4.6 shows that the computed F-statistic for Data M (the Initial Dataset) and Data N (the Initial Dataset) are 12.119 and 119.701, respectively, (P-values equivalent of about 0.000) led to the conclusion that both models are of good-fit to Data M (the Initial Dataset) and Data N (the Initial Dataset).

As shown in Table 4.5, the R-square value of 0.538 indicates that approximately 53.80% of the total variability in the dependent variable, Y_M, is attributable to the fluctuations being in the independent variables, X_{1M}, X_{2M}, X_{3M}, X_{4M}, and X_{5M}; while approximately 46.20% of the variation remains unexplained, potentially due to omitted variables. Table 4.6 also reveals that the R-square value of 0.865 indicates approximately 86.50% of the total variability in the dependent variable, Y_N, is attributable to the fluctuations being in the independent variables, X_{1N}, X_{2N}, and X_{3N}; while approximately 13.50% of the variation remains unexplained, potentially due to omitted variables. The ancillary statistics for the regression model (in Table 4.7) for Data M (the Initial Dataset), the values of AIC, SBC and RMSE are 1071.424, 1083.787 and 9774.849; while for Data N (the Initial Dataset), the values of AIC, SBC and RMSE are 768.443, 776.820 and 584.946.

Although other assumptions such as multicollinearity and autocorrelation were met, the failure of normality and constant variance assumptions in Data M (the Initial Dataset) and Data N (the Initial Dataset) necessitated remediation to address non-normality and nonconstant variance issues. The correction is done by employing the semi-logarithm transformation method for Data M (the Initial Dataset) and Data N (the Initial Dataset) (see Nwankwo, 2011) as expressed in (9); which in this case are given by,

$$InY_M^{\frac{1}{2}} = \beta_{0M} + \beta_{1M} InX_{1M}^{\frac{1}{2}} + \beta_{2M} InX_{2M}^{\frac{1}{2}} + \beta_{3M} InX_{3M}^{\frac{1}{2}} + \beta_{4M} InX_{4M}^{\frac{1}{2}} + \beta_{5M} InX_{5M}^{\frac{1}{2}} + u_M \quad (14)$$

And

$$InY_N^{\frac{1}{2}} = \beta_{0N} + \beta_{1N} InX_{1N}^{\frac{1}{2}} + \beta_{2N} InX_{2N}^{\frac{1}{2}} + \beta_{3N} InX_{3N}^{\frac{1}{2}} + u_M \quad (15)$$

The procedure of the Multiple Linear Regression Analysis is carried out on Data M (the Initial Dataset) and Data N (the Initial Dataset) which failed the normality and constant variance assumptions but is now corrected. The results outputs of the procedure are presented in Tables 4.8 to 4.14 and Figure 4.3 to 4.4.

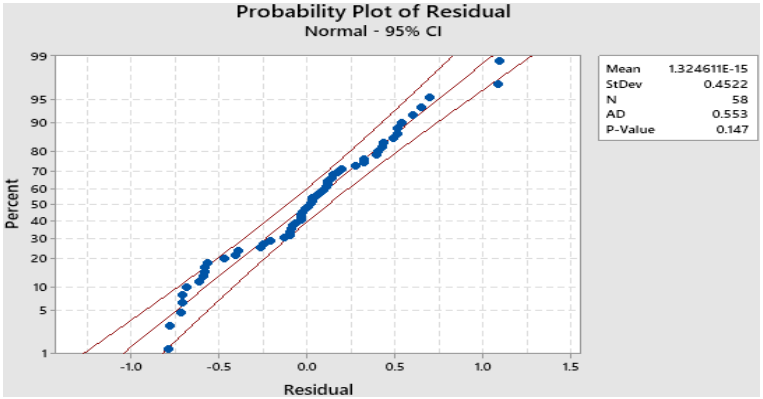

Fig. 4.3: The Anderson-Darling Test for the Normality Assumption on Data M (Non-normality and Nonconstant Variance Corrected).

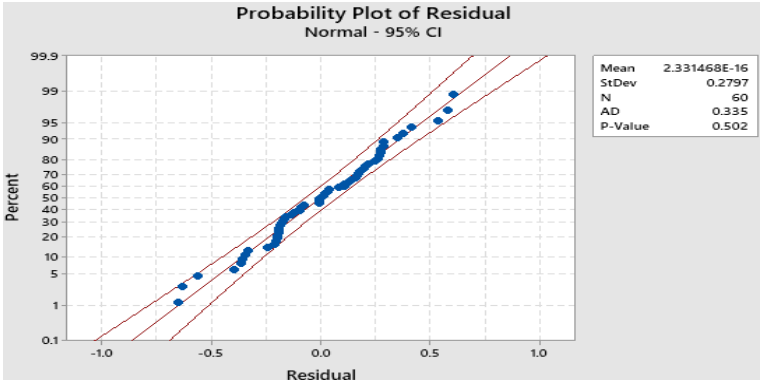

Fig. 4.4: The Anderson-Darling Test for the Normality Assumption on Data N (Non-normality and Nonconstant Variance Corrected).

Table 4.8: D'Agostino Omnibus Test for Normality Assumption on Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected)

	Data M (Non-normality and Nonconstant Variance Corrected)	Data N (Non-normality and Nonconstant Variance Corrected)
K^2-stat	0.214	0.065
p-value	0.8987	0.9681
alpha	0.05	0.05
normal	yes	yes

Table 4.9: White Test for Constant Variance Assumption on Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected)

	Data M (Non-normality and Nonconstant Variance Corrected)	Data N (Non-normality and Nonconstant Variance Corrected)
LM stat	0.538	2.957
df	2	2
p-value	0.7641	0.2280
F stat	0.257	1.4773
df1	2	2

df2	55	57
p-value	0.7740	0.2369

Table 4.10: Regression Model Coefficients for Data M (Non-normality and Nonconstant Variance Corrected)

Variable	Unstandardized Coefficients		Standardized Coefficients	t-stat	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
β_{0M}	0.972	0.958		1.014	0.315					
X_{1M}	0.396	0.113	0.543	3.503	0.001	0.762	0.437	0.244	0.202	4.942
X_{2M}	0.213	0.104	0.347	2.057	0.045	0.671	0.274	0.143	0.171	5.859
X_{3M}	0.021	0.155	0.022	0.138	0.891	0.192	0.019	0.010	0.196	5.112
X_{4M}	0.049	0.031	0.152	1.585	0.119	0.652	0.215	0.110	0.531	1.884
X_{5M}	-0.022	0.631	-0.003	-0.035	0.972	0.526	-0.005	-0.002	0.498	2.007

Table 4.11: Regression Model Coefficients for Data N (Non-normality and Nonconstant Variance Corrected)

Variable	Unstandardized Coefficients		Standardized Coefficients	t-stat	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
β_{0N}	0.336	0.881		0.382	0.704					
X_{1N}	1.077	0.143	1.013	7.551	0.000	0.933	0.710	0.363	0.128	7.809
X_{2N}	-0.093	0.138	-0.089	-0.672	0.504	0.857	-0.089	-0.032	0.131	7.659
X_{3N}	0.064	0.249	0.013	0.257	0.798	0.217	0.034	0.012	0.944	1.059

Table 4.12: Model Summary for Data M (Non-normality and Nonconstant Variance Corrected)

Multiple R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
				R Square Change	F Change	df1	df2	Sig. F Change	
0.865	0.748	0.723	0.4734764	0.748	30.795	5	52	0.000	1.828

Table 4.13: Model Summary for Data N (Non-normality and Nonconstant Variance Corrected)

Multiple R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
				R Square Change	F Change	df1	df2	Sig. F Change	
0.933	0.871	0.864	0.28706	0.871	125.934	3	56	0.000	1.822

Table 4.14: Additional Overall Fit of the Regression Models for Data M (Non-normality and Nonconstant Variance Corrected)

	Data M (Non-normality and Nonconstant Variance Corrected)	Data N (Non-normality and Nonconstant Variance Corrected)
AIC	-81.061	-145.907
SBC	-68.699	-137.529
RMSE	0.473	0.287
Ave. Abs. PCT Error	9.576	9.066

Given the insignificant results from the normality and constant variance tests, going forward to transform Data M (the Initial Dataset) and Data N (the Initial Dataset) using a semi-logarithm approach. The tests for normality and constant variance will be re-conduct to verify whether the transformations have successfully corrected the existence of non-normality and nonconstant variance of residuals. From the normal probability plot and the Anderson-Darling (AD) test in Figs. 3 and 4, which were used to verify that the residuals are normally distributed, the computed Anderson-Darling statistic, AD, for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) are 0.553 and 0.335, respectively, with (P=0.147 and 0.502, respectively); which are greater than the level of significance, $\alpha = 0.05$. Therefore, the null hypothesis (which states that residuals are normally distributed) is not rejected; thus, the conclusion is that the residuals are now normally distributed. Also giving support to this conclusion are the D'Agostino-Omnibus test (in Table 4.8) for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) which are 0.214 and 0.065, respectively, with (P=0.8987 and 0.9681, respectively); which are greater than the level of significance, $\alpha = 0.05$.

White test in Table 4.9 was used to test for the constant variance assumption, the value of the computed test statistic for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) are 0.538 and 2.957, respectively, with (P=0.7641 and 0.2280, respectively); which are greater than the level of significance, $\alpha = 0.05$. Therefore, the null hypothesis (which states that the residuals are homoscedastic) is not rejected; thus, the conclusion is that the residuals are now homoscedastic. The multiple linear regression is now carried out on Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected), and the results outputs are as presented from Tables 4.10 to 4.11, the multiple linear regression models for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) are obtained as,

$$\widehat{InY}_M^{\frac{1}{2}} = 0.972 + 0.396InX_{1M}^{\frac{1}{2}} + 0.213InX_{2M}^{\frac{1}{2}} + 0.021InX_{3M}^{\frac{1}{2}} + 0.049InX_{4M}^{\frac{1}{2}} - 0.022InX_{5M}^{\frac{1}{2}} + u_M \quad (14)$$

And

$$\hat{InY}_{N}^{\frac{1}{2}} = 0.336 + 1.077InX_{1N}^{\frac{1}{2}} - 0.093InX_{2N}^{\frac{1}{2}} + 0.064InX_{3N}^{\frac{1}{2}} + u_{M} \quad (15)$$

Also shown in Table 4.10, the VIF values < 10 for all five independent variables in Data M (Non-normality and Nonconstant Variance Corrected), indicating no multicollinearity issues; and Table 4.11 shows VIF values < 10 for all three independent variables in Data N (Non-normality and Nonconstant Variance Corrected), indicating no multicollinearity issues. From Tables 4.12 and 4.13, the value of the computed Durbin-Watson statistic for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) are 1.8 and 1.8, respectively, (which are approximately equal to 2) implies that there is no existence of autocorrelation in the models. Furthermore, Tables 4.12 and 4.13 shows that the computed F-statistic for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) are 30.795 and 125.934, respectively, (P-values equivalent of about 0.000) led to the conclusion that both models are of good-fit to Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected).

As shown in Table 4.12, the R-square value of 0.748 indicates that approximately 74.80% of the total variability in the dependent variable, $InY_{M}^{\frac{1}{2}}$, is attributable to the fluctuations being in the independent variables, $InX_{1M}^{\frac{1}{2}}, InX_{2M}^{\frac{1}{2}}, InX_{3M}^{\frac{1}{2}}, InX_{4M}^{\frac{1}{2}}$ and $InX_{5M}^{\frac{1}{2}}$; while approximately 25.20% of the variation remains unexplained, potentially due to omitted variables. Table 4.13 also reveals that the R-square value of 0.871 indicates approximately 87.10% of the total variability in the dependent variable, $InY_{N}^{\frac{1}{2}}$, is attributable to the fluctuations being in the independent variables, $InX_{1N}^{\frac{1}{2}}, InX_{2N}^{\frac{1}{2}}$, and $InX_{3N}^{\frac{1}{2}}$; while approximately 12.90% of the variation remains unexplained, potentially due to omitted variables. The ancillary statistics for the regression model (in Table 4.14) for Data M (Non-normality and Nonconstant Variance Corrected), the values of AIC, SBC and RMSE are -81.061, -61.699 and 0.473; while for Data N (Non-normality and Nonconstant Variance Corrected), the values of AIC, SBC and RMSE are -145.907, -137.529 and 0.287.

5. Conclusion

This study demonstrates the significance of correcting for Non-normality and Nonconstant variance of residuals in linear regression modelling. To showcase this concept, this study utilized two hypothetical datasets; these datasets were labelled Data M (the Initial Dataset) and Data N (the Initial Dataset). Both initial datasets satisfied the multicollinearity, and autocorrelation assumptions but failed to meet the assumptions of normality and constant variance, indicating the existence of non-normality and nonconstant variance of residuals. The method adopted in this study to correct for non-normality and nonconstant variance is not limited to regression analysis, but can be extended to other parametric tests, such as Analysis of Variance (ANOVA), when these assumptions are violated in other parametric analyses.

The ordinary least square method was employed to estimate the multiple linear regression models for Data M (the Initial Dataset) and Data N (the Initial Dataset), respectively. The model established for Data M (the Initial Dataset) is statistically significant (indicating a good fit to the dataset) with an R-square value of 0.538, an AIC value of 1071.424, an SBC value of 1083.787, and an RMSE value of 9774.849. Correspondingly, the model established for Data N (the Initial Dataset) is statistically significant (indicating a good fit to the dataset) with an R-square value of 0.865, an AIC value of 768.443, an SBC value of 776.427, an RMSE value of 581.946.

To alleviate the issues of non-normality and nonconstant variance of residuals in the two initial datasets, semi-logarithm transformation method was employed on the variables in Data M (the Initial Dataset) and Data N (the Initial Dataset). These transformations gave rise to new sets of data now referred to as, Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected). Following the correction, the datasets met all the linear assumptions required for regression analysis.

The model established for Data M (Non-normality and Nonconstant Variance Corrected) is statistically significant (indicating a good fit to the dataset) with an R-square value of 0.748, an AIC value of -81.061, an SBC value of -61.699, and an RMSE value of 0.473. In a similar vein, the model established for Data N (Non-normality and Nonconstant Variance Corrected) is statistically significant (indicating a good fit to the dataset) with an R-square value of 0.871, an AIC value of -145.907, an SBC value of -137.529, an RMSE value of 0.287. On a lighter note, comparing the models using R-square, AIC, SBC, and RMSE values of the original and transformed datasets reveals interesting insights. Notably, significant improvements were observed in the transformed models compared to the original models. The values of the R-square for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) are, respectively, greater than the values of the R-square for Data M (the Initial Dataset) and Data (the Initial Dataset). It is noticeable that 0.748>0.538 and 0.871>0.865. Also, the value of the AIC for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) are, respectively, less than the values of the AIC for Data M (the Initial Dataset) and Data N (the Initial Dataset). It is noticeable that -81.061<1071.424 and -145.907<768.443. The trends observed in the AIC comparisons are also reflected in the SBC and RMSE comparisons of the models.

From the perspective of R-square, AIC, SBC and RMSE values, it will be concluded that the estimated regression models for Data M (Non-normality and Nonconstant Variance Corrected) and Data N (Non-normality and Nonconstant Variance Corrected) demonstrate superior model performance when compared to the regression models for Data M (the Initial Dataset) and Data N (the Initial Dataset).

References

[1] Barker, L. E. and Shaw, K. M. (2015). Best (but oft-forgotten) Practices: Checking Assumptions Concerning Regression Residuals. Am J Clin Nutr 102:533–9. https://doi.org/10.3945/ajcn.115.113498.

[2] Bartlett, M. S. (1947). The use of transformation. Biometric Bulletin, 3, 39-52. https://doi.org/10.2307/3001536.

[3] D'Agostino, Ralph B., Albert Belanger, Ralph B., and D'Agostino, Jr. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. The American Statistician, 44(4): 316–321. https://doi.org/10.1080/00031305.1990.10475751.

[4] Das, K. R. and Imon, A. H. M. R. A Brief Review of Tests for Normality. American Journal of Theoretical and Applied Statistics. 5(1): 5-12. https://doi.org/10.11648/j.ajtas.20160501.12.

[5] Flachaire, E. (2005). Bootstrapping Heteroskedastic Regression Models: Wild Bootstrap vs Pairs Bootstrap. Computational Statistics and Data Analysis, 49 (2): 361-376. https://doi.org/10.1016/j.csda.2004.05.018.

[6] Gujarati, D. (2004). Basic econometrics (4th ed.). Mcgraw-Hill, New York, U.S.A.

[7] Hawkins, D. L. (1989). Using U Statistics to Derive the Asymptotic Distribution of Fisher's Z Statistic. American Statistician, 43, 235-237. https://doi.org/10.1080/00031305.1989.10475666.

[8] Hogg, R. V. (1979). An Introduction to Robust Estimation, in Robustness in Statistics, Edited by Launer, R. L., Wilkinson, G. N., New York: Academic Press; 1–17. https://doi.org/10.1016/B978-0-12-438150-6.50007-8.

[9] Jude, O. and Isobeye, G. (2021). Effect of Non-Normal Error Distribution on Simple Linear/Non-Parametric Regression Models. International Journal of Statistics and Applied Mathematics, 6(4): 131-136.
[10] Judge, G. G., Griffith, W. E., Hill, R. C., Lutkepohl, H., and Lee, T. (1985). Theory and Practice of Econometrics. (2nd ed.). John Wiley and Sons, New York, USA.
[11] Kim, T. K. and Park, J. H. (2009). More about the basic assumptions of t-test: normality and sample size. The Korean Society of Anesthesiologists, 72(4): 331-335. https://doi.org/10.4097/kja.d.18.00292.
[12] Koenker, R. W. (1982). Robust Methods in Econometrics. Econometric Reviews 1: 213-290. https://doi.org/10.1080/07311768208800017.
[13] Koutsoyiannis, A. (1977). Theory of econometrics (7th ed.). Macmillian, London, United Kingdom. https://doi.org/10.1007/978-1-349-09546-9.
[14] Kutner, M. H., Nacthtsheim, C. J., Neter, J. and Li Williams (2005). Applied linear statistical models (5th ed.). Mcgraw-Hill/Irwin, New York, U.S.A.
[15] Neave, H. R. (1978). Statistics Tables for Mathematics, Engineers, Economics and the Behavioural and Management Sciences. George Allen and Unwin, London, United Kingdom.
[16] Nwankwo, S. C. (2011). Econometrics: a practical approach. El'demak, Enugu, Nigeria.
[17] Osemeke, R. F., Igabari, J. N. and Nwabenu, D. C. (2024). Detection and Correction of Violations of Linear Model Assumptions by Means of Residuals. Journal of Science Innovation & Technology Research 3 (9): 1-15
[18] Ohaegbulem, E. U. and Iheaka, V. C. (2024). On Remedying the Presence of Heteroscedasticty in a Multiple Linear Regression Modelling. African Journal of Mathematics and Statistics Studies, 7(2): 225-261.
[19] Osaro, A. D. (2023). Application of Transformation of Variables in Remedying Heteroscedasticity in Nigeria GDP, Conditioning and Some Fiscal Variables. NIPES Journal of Science and Technology Research 5(1): 84-91.
[20] Pedace, R. (2013). Econometrics for Dummies. John Wiley & Sons, New Jersey, Canada.
[21] Thinh, R., Samarta, K. and Jansakula, N. (2020). Linear Regression Models for Heteroscedastic and Non-Normal Data. ScienceAsia 46: 353-360 https://doi.org/10.2306/scienceasia1513-1874.2020.047.
[22] White, H. (1980). A Heteroskedastic Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity. Open Access Library Journal, 4(8):817-818. https://doi.org/10.2307/1912934.

Appendix A

Table 3.1: The Hypothetical Data M (the Initial Dataset)

Y_M	X_{1M}	X_{2M}	X_{3M}	X_{4M}	X_{5M}
121.6	5,281.10	0.7143	13.76	0.00001	0.3109
319.6	6,650.90	0.6955	16	0.00002	0.3567
248.3	7,187.50	0.6579	3.46	0.00001	0.3373
192.6	8,630.50	0.6579	5.4	0.00002	0.4059
48.3	18,823.10	0.6299	12.67	0.00001	0.4002
475.4	21,475.20	0.6159	33.96	0.00002	0.4027
46.3	26,655.80	0.6265	24.3	0.00001	0.4464
197.6	31,520.30	0.6466	15.09	0.00001	0.4671
331.8	34,540.10	0.606	21.71	0.00001	0.4133
289.9	41,974.70	0.5957	11.7	0.00001	0.4362
467	49,632.30	0.5464	9.97	42	0.4691
137.3	47,619.70	0.61	20.9	30.2	0.5011
1,624.90	49,069.30	0.6729	7.7	29.2	0.3867
556.7	53,107.40	0.7241	23.2	35.7	0.3089
534.8	59,622.50	0.7649	39.6	47.5	0.2728
329.7	67,908.60	0.8938	5.5	47	0.2766
2499.6	69147	2.0206	5.4	31.8	0.215544
680	105222.8	4.0179	10.2	33.4	0.458288
1345	139085.3	4.5367	38.3	29.2	0.378462
439.4	216797.5	7.3916	40.9	40.5	0.409744
464.3	267550	8.0378	7.5	47.5	0.581588
1808	312139.7	9.9095	13	42.5	0.676055
8269.2	532613.8	17.2984	44.5	35.7	0.654814
32994.4	683869.8	22.0511	57.2	48.5	0.562095
3907.2	899863.2	21.8861	57	41.1	0.409893
48677	1933212	21.8861	72.8	38	0.88236
2731	2702719	21.8861	29.3	40.1	0.692699
5730.9	2801973	21.8861	8.5	39.4	0.744968
24078.8	2708431	21.8861	10	26	0.586788
1779.1	3194015	92.6934	6.6	32.6	0.642291
3347	4582127	102.1052	6.9	46.9	0.639604
3377	4725086	111.9433	18.9	39.9	0.682767
8205.5	6912381	120.9702	12.9	28	0.471165
13056.5	8487032	129.3565	14	34.4	0.608943
19909.1	11411067	133.5004	15	37.4	0.577494
25881.8	14572239	132.147	17.9	43.1	0.689488
41470.8	18564595	128.6516	8.2	38.1	0.561994
54041.9	20657318	125.8331	5.4	34.8	0.591641
49456.2	24296329	118.5669	11.58	37	0.631836
41429.4	24794239	148.9017	11.54	25.5	0.542824
9073.04	29205783	150.298	13.72	32.6	0.651966
6121.6	43012.51	191.2	147.57	7.75	0.5109
7319.6	54612.26	160.9	157.18	10.25	0.3567
2152	62980.4	248.8	164.21	10	0.3373
2757.4	71713.94	337.2	185.98	12.5	0.4059
2954.4	80092.56	428.2	196.17	9.25	0.4002
3259.6	89043.62	487.1	242.26	10.5	0.4027
4677.3	94144.96	947.7	312.5	17.5	0.4464
4227.8	101489.5	701.1	410.77	16.5	0.4671
4991.3	113711.6	817.5	489.77	26.8	0.4133

5349	127736.8	1018.2	584.25	25.5	0.4362
1387.33	144210.5	1226	897.12	20.01	0.4691
1631.5	152324.1	1504.2	1244.8	29.8	0.5011
2111.51	49,069.30	1919.7	1751.28	18.32	0.3867
3478.91	53,107.40	4038	369.43	21	0.4089
5787.51	59,622.50	2450.9	4045.32	20.18	0.3728
7759.2	67,908.60	3240.8	4374.5	19.74	0.6766
12705.62	21169147	3453	510.4	13.54	0.54697

Appendix B

Table 3.2: The Hypothetical Data N (the Initial Dataset)

Y_N	X_{1N}	X_{2N}	X_{3N}
11.5	5	58	850
200	138	2454	954
70	44	573	874
100	98	2172	941
7000	1651	31123	1185
70	26	295	874
125	35	1131	902
2200	1278	22571	1048
400	365	6554	960
110	47	793	930
6000	1650	36330	1142
58.4	39	522	800
212	69	1041	1060
400	57	1059	1000
1888	896	16411	1150
486	125	1678	1170
439	135	2529	110
1900	653	19082	1080
155	114	3523	1026
6.9	11	207	873
509	346	6781	1097
180	25	147	990
53	21	214	920
100	44	764	900
30	18	176	1176
157	99	3682	930
475	223	5665	1037
613	384	4411	960
483	141	3341	860
2500	2021	4528	1086
142	66	1251	1030
210	73	1036	1000
20	10	120	1070
150	94	2344	858
300	195	2400	1180
233.5	70	1416	910
235	165	4148	1001
460.7	316	9738	980
1632	355	5578	1060
93	30	505	930
263	185	3724	1124
144.5	101	2387	945
770	148	1900	1190
1700	960	16750	1057
1100	284	2833	1310
1900	905	15762	1090
60	55	875	848
1200	445	6603	1060
1600	623	14727	1120
1289	412	11179	1230
1666	1607	9251	883
15	26	608	800
160	48	656	1010
200	281	3892	980
263	195	2987	1070
487	275	5148	1060
3300	867	1240	1260
145	37	569	843
205	28	628	980
7377	2606	34055	1160