International Journal of Advanced Statistics and Probability, 11 (2) (2024) 67-83

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP

Research paper

On remedying the presence of heteroscedasticity in a multiple linear regression modelling

Emmanuel Uchenna Ohaegbulem [1] *, Victor Chijindu Iheaka [1]

[1] *Department of Statistics, Imo State University, Owerri, Imo State, Nigeria*
**Corresponding author E-mail: emmanx2002@yahoo.com*

Abstract

This study demonstrated the very essence of remedying the presence of heteroscedasticity, where it existed, in regression modelling. Two different hypothetical data, Data A (the Original) and Data B (the Original), were used in this study for the purpose of illustration. The normality, multicollinearity and autocorrelation assumptions were satisfied, but the Breusch-Pagan test and the White test established the existence of heteroscedasticity in the two datasets. The estimated multiple linear regression model for Data A (the Original) was statistically significant with an R-square value of 0.976, an AIC value of 332.5929, and an SBC value of 347.2533; and the one for Data B (the Original) was also statistically significant with an R-square value of 0.553, an AIC value of 69.89669, and an SBC value of 82.15499. The Log-transformation was applied on the variables in Data A (the Original) and Data B (the Original) to give rise to new sets of data, Data A (Now with Heteroscedasticity Remedied) and Data B (Now with Heteroscedasticity Remedied); which equally satisfied the normality, multicollinearity and autocorrelation assumptions, and also satisfied that there were no existence of heteroscedasticity in the two datasets. Now, the estimated multiple linear regression model for Data A (Now with Heteroscedasticity Remedied) was statistically significant with an R-square value of 0.986, an AIC value of -135.021, and an SBC value of -120.361; and the estimated model for Data B (Now with Heteroscedasticity Remedied) was statistically significant with an R-square value of 0.624, an AIC value of -32.0801, and an SBC value of -19.8218. From the points of view of the values of the R-square (0.986>0.976 and 0.624>0.553), AIC (-135.021<332.5929 and -32.0801<69.89669) and SBC (-120.361<347.2533 and -19.8218<82.15499), it was evident that the estimated regression models for Data A (Now with Heteroscedasticity Remedied) and Data B (Now with Heteroscedasticity Remedied) were, respectively, better models when compared to the regression models for Data A (the Original) and Data B (the Original).

Keywords: *Multiple Linear Regression Analysis; Correlation Analysis; Heteroscedasticity; Autocorrelation; Multicollinearity; Remedying.*

1. Introduction

Regression analysis is a set of statistical processes for establishing the relationship among related variables. According to Gujarati [7], regression analysis is concerned with the study of the dependence of one variable (the dependent variable) on one or more other variables (the independent variables) with a view to estimating or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter. The Ordinary Least Squares (OLS) method is one of the statistical tools widely used to estimate the parameters of the linear regression model. Under the usual assumptions, the least-squares estimators possess many desirable properties.

Virtually every introduction to OLS regression includes an overview of the assumptions behind this method to make sure that the inferences obtained from it are warranted [2]. Homoscedasticity is one of the most important assumptions of the OLS method. Homoscedasticity simply implies that the error terms for each observation are the same for all observations. However, in a situation where the error terms do not have constant variance, it is regarded to be heteroscedastic. Heteroscedasticity is usually defined as some variation of the phrase " non-constant error variance" , or the idea that, once the predictors have been included in the regression model, the remaining residual variability changes as a function of something that is not in the model [3 - 5], [12].

There are several causes of heteroscedasticity, one of which is incorrect functional form of the regression model. Williams [18] opined that measurement error can cause heteroscedasticity, and also a situation where there are subpopulation differences or other interaction effects. Heteroscedasticity or unequal variances, often occurs in cross-sectional data; for instance, mixing datasets with different measures of scale. Clearly, regression models with cross sectional data, especially in cases where the scale of the dependent variable varies across observations, heteroscedasticity is more likely to occur. According to Gujarati [7], heteroscedasticity can also arise as a result of the presence of outliers (that is, observation from a different population to that generating the remaining sample observation). Another cause of heteroscedasticity is model misspecification. Gujarati [7] explained that heteroscedasticity may be present in the regression model due to the fact that some important variables are omitted from the model. According to Nwakuya and Nwabueze [13], most economic data show the presence of heteroscedasticity; and heteroscedasticity mostly occurs because of underlying errors in variables, outliers, misspecification of model amongst others.

The problem of heteroscedasticity imposes a great challenge for estimation of the regression model. In the presence of heteroscedasticity, the OLS estimators and the forecasts based on them would still be unbiased and consistent, but they would no longer be BLUE. According to Adepoju and Tayo [1], the most damaging consequence of heteroscedasticity is that the OLS estimators of the parameter covariance matrix, whose diagonal elements are used to estimate the standard errors of the regression coefficients, become biased and inconsistent. The effects of heteroscedasticity can be severe, as it can result to the estimates of the regression coefficients being biased and inconsistent; which can have serious consequences for hypothesis testing, decision-making, and also reduce the statistical power of the analysis. Hayes and Cai [9] explained that the outcome of the test statistic from the regression model is not influenced by heteroscedasticity either, but the F-test and t-test associated are being affected.

Consequently, lack of efficiency of the OLS estimators due to the presence of heteroscedasticity makes the forecasting and prediction from the model to be unreliable. Therefore, a remedial measure is surely to be called for. Remedying the presence of heteroscedasticity in the regression model will guarantee making the OLS estimators of the regression model parameters to be more reliable. In other words, the remediation of heteroscedasticity in the regression model is paramount in order to obtain the estimators that are BLUE. However, when there is no presence of heteroscedasticity, one will simply go ahead with the regression modelling; as have been shown in the literature (see, for example, Ohaegbulem and Iheaka, [15]). Some past works have also showcased the fact that the presence of heteroscedasticity in regression modelling had to be remedied before going on to arrive at better models with valid and more reliable estimates for further inferences. A few of these past works are reviewed here.

Nwakuya and Nwabueze [13] employed the Ordinary Least Square (OLS) regression to establish the Multiple Linear Regression (MLR) model of the relationship that existed among GDP (Y) and Inflation (X_1), Trade-index (X_2), Civil-liability (X_3) and Population (X_4) in an economic data called Africa (collected from six African countries with a sample size of 120 for each variable), which they said they got from the R package. They applied five different heteroscedasticity tests (which include Park test, Glejser test, Goldfeld Quandt test, White test and Breusch-Pagan test) and all the tests showed presence of heteroscedasticity, as they confirmed statistical significance at 5% level of significance. The result of the estimated multiple regression model before the application of the Box-Cox transformation as a corrective measure to the presence of heteroscedasticity was given as,

$$\hat{Y} = 7.41 - 6.635X_{1i} + 19.50X_{2i} - 638.4X_{3i} + 0.001812X_{4i}.$$

The results before Box-Cox transformation also revealed that an AIC and SIC values were obtained as 1667.924 and 1684.394, respectively. The p-Value of 0.000 (or equivalently, an F-statistic of 63.96) showed that the regression model was significant (implying that the regression model was of good-fit to the dataset). The Box-Cox transformed data proved to be normally distributed, with a p-Value of 0.057. Also, it was confirmed that there was no multicollinearity among the regressors. The result of the regression estimated model after Box-Cox transformation was given as,

$$\hat{Y} = 54.40 + 0.000295X_{1i} - 0.001038X_{2i} + 0.02223X_{3i} - 0.000000002934X_{4i}.$$

The results after Box-Cox transformation also revealed that an AIC and SIC values were obtained as -640.6783 and -624.2087, respectively. An R^2 values before the transformation and after the transformation were obtained as 0.6993 and 0.7341, respectively. The p-Value of 0.000 (or equivalently, an F-statistic of 75.94) showed that the model was significant. The result of the Park test, Glejser test, Goldfeld Quandt test and Breusch Pagan test confirmed statistical insignificance at 5% level of significance (with p-Values of 0.3397, 0.2968, 0.9838, 0.2009, respectively); and also the White test was also insignificant $(nR^2 = 1.053 < \chi_9^2 = 3.325)$. In conclusion, the values of the R^2 and AIC had demonstrated that the model after the transformation was a better regression model compared to the regression model before the transformation.

Zhou et al. [20] carried out a multiple regression model among the dependent variable, logBaseCr, and the explanatory variables, Age, Gender, logWeight, Albumin and Haemoglobin. The result showed that the multiple regression model among logBaseCr and the explanatory variables was,

$$logBaseCr = 0.0070Age + 0.1130Gender + 0.2578logWeight + 0.0755Albumin - 0.0421Hgb$$

The results also showed that Age, logWeight and Haemoglobin had a statistically significant relationship with logBaseCr at 5% level of significance (with p-Values of 0.01, 0.019, 0.01, respectively), while Gender and Albumin were not statistically significant factors. The AIC and SIC values were obtained as 207.1 and 229.0, respectively. The proposed two-step procedure was applied to examine the patterns of residual plots rigorously. The covariate specific p-Values were found to be 0.086, 0.25, 0.25, 0.0004 and 0.93 for Age, Gender, log-Weight, Albumin and Hgb, respectively. Given the significance level 0.05, the null hypothesis of homoscedasticity was rejected. The Weighted Least Squares (WLS) method was employed to correct the presence of heteroscedasticity in the regression model. With the correction, the multiple regression model among logBaseCr and the explanatory variables was given as,

$$logBaseCr = 0.0540Age + 0.1398Gender + 0.2790logWeight + 0.0889Albumin - 0.0466Hgb$$

The results also revealed that all the regressors used in this study had a statistically significant relationship with logBaseCr log at 5% level of significance (with p-Values of 0.01, 0.036, 0.01, 0.041 and 0.010, respectively). The AIC and SIC values were obtained as 185.2 and 207.1, respectively. It was concluded that misspecification of the random effects structures may affect the estimation efficiency of the fixed effects.

Gidigbi and Donga [6] studied the domestic, foreign direct investment and economic growth nexus in selected African countries. Multiple regression analysis was used to analyse the logarithm of data on GDP (LGDP), Gross Domestic Investment (LGDI), Foreign Direct Investment (LFDI) and Current Account Balance (LCAB). The unit root test results revealed that LGDP was stationary at first difference at 1% statistical significance level with the value of statistic, -16.5227. In addition, LGDI, LFDI and LCAB were stationary at first difference as indicated by the Levin, Lin & Chu t* statistic values of -121.580, -17.2090 and - 4.60164, which was statistically significant at 1% significance level, respectively. The results of the Cointegration test revealed that LGDP, LGDI, LFDI, and LCAB, exhibited a long-run relationship, which implied that the variables could be put together in a regression model. Panel regression estimation was adopted in correcting the problem of heteroscedasticity and autocorrelation estimation. The results also revealed that all the regressors had a statistical

significant relationship with LGDP at 1% and 5% level of significance. An R^2 value of 0.93 was obtained, which implied that the regressors accounted for almost 93.20% of the total variations in the regressand. This indicated that about 6.80% variability could be attributed to other regressors outside the ones featured in the model. The F-statistic of 3443.27 implied that the model was good fit to the dataset at 1% significance level. It was concluded that the investment in general and domestic investment, in particular, was very relevant to the economic growth in the continent.

Jabłońska [10] conducted a study involving the modelling of the quality of life of older people. The Multiple regression analysis was used to explain the effect of the regressors on the quality of life for both men and women ranging from the age of sixty and above. The results of quality of life model (both men and women) assumptions indicated that errors normality with mean equals 0, significant linear structure of the model, no autocorrelation and no multicollinearity. Also, the results of testing the homoscedasticity assumptions in the model yielded White s test (p = 0.001) and BPG test (p = 0.009) for Men' s quality of life model, verified that the test was insignificant. Also, testing the homoscedasticity assumption in Women' s quality of life model showed that the test was significant (White test: p=0.452; BPG test: p=0.590). Thus, the result showed that the model among Men' s quality of life and the regressors was,

$$QL = -0.095BML - 0.057Age - 1.169ADL + 0.189SN - 0.114LO + 0.58PLC + 1.654PR + 0.15SS$$

Heteroscedasticity-Consistent covariance matrix estimators (HC-estimators) were used to correct for the presence of heteroscedasticity in the model. The result showed HC4m was the best for the model because it was much more conservative than HC1, HC2 and HC3. As a result of the use of HC4m-estimator, four variables (ADL, social network, loneliness and social support) were considered significant in the context of men' s quality of life. The result also showed the relationship between Women' s quality of life and the regressors was,

$$QL = -0.01BML - 0.114Age - 1.121ADL + 0.128SN - 0.076O + 1.274PLC + 0.887PR + 0.122SS$$

The result demonstrated that OLS method had similar result with HC2m, and with the use of HC4m, four variables (ADL, SN, loneliness and social support) were considered significant in the context of Women' s quality of life. It was concluded that the use of HC4m was preferable for correcting the presence of heteroscedasticity in the model.

Thus, this present study is centred on expressing the very essence of going about the remediation of the presence of heteroscedasticity (where/when it occurs) in a regression model; and not to (as is usually the case with most random researchers) go ahead with the estimation of the regression model parameters and the onward engagement of making predictions with the model so established without correcting for the presence of heteroscedasticity.

2. Method

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. This is achievable with the use of the Ordinary Least Squares (OLS) procedure to estimate the coefficients for the independent variables. Furthermore, the analysis shall evaluate the contributions of each of the independent variables to the dependent variable.

The multiple linear regression model, which explains the relationship that exists among the dependent and independent variables, is usually given as,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e_i \tag{2.1}$$

Where,

Y, X' s, β' s and e_i (i = 1, 2, ···, k) are the dependent variable, independent variables, the estimated parameters, and the error term, respectively.

Equation (2.1) can also be expressed in matrix terms (see, for example Kurtner et al. [12]) as,

$$\underset{(n\times 1)}{Y} = \underset{(n\times k)}{X}\underset{(k\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon} \tag{2.2}$$

Where,

$$Y' = (Y_1, Y_2, \cdots, Y_n) \tag{2.3}$$

$$X = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{k1} \\ X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix} \tag{2.4}$$

$$\beta' = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k) \tag{2.5}$$

And

$$\varepsilon' = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) \tag{2.6}$$

Applying the Ordinary Least Squares (OLS) method (see, for example Kurtner et al. [12]) the regression model parameters, β_i ' s, are estimated as,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.7}$$

Then, the estimated regression model will be obtained by substituting the values of the $\hat{\beta}_i$ ' s in (2.7) into (2.1).

Once the multiple regression model is developed, its predictive accuracy would be evaluated using the coefficient of multiple determination, R^2, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) or the Schwarz-Bayesian Criterion (SBC).

Prior to the estimation of the multiple regression model expressed above, a number of assumptions are to be fulfilled so that the estimated regression model parameters will be reliable. According to Gujarati [7], these assumptions include the followings;

i) The regression model must be linear in the parameters.

ii) The independent variable, X, is assumed to be non-stochastic. That is, the values taken by the independent variables, X, are considered fixed in repeated samples.

iii) The error terms, e_i, are normally distributed, having the expected value or mean of zero.

iv) Homoscedasticity or equal variance of the error terms, e_i. That is, the variance of e_i is the same for the observations.

v) No serial correlation or zero autocorrelation between the disturbances. Given that any two X values, X_i and X_j $(i \neq j)$, the correlation between any two error e_i and e_j is zero.

vi) The covariance between e_i and X_i is zero. That is, $E(e_i \mid X_i) = 0$.

vii) The independent variables are linearly independent (that is, it is not possible to express any independent variable as a linear combination of the other). In other words, there is no perfect multicollinearity.

viii) The number of observations must be greater than the number of parameters to be estimated.

ix) Variability in the X_i values exists (that is, the X_i values in a given sample must not all be the same).

x) The regression model is correctly specified bias and the independent variables are measured with no error.

Interestingly, though, the most pronounced assumptions that are supposedly to be met are those of Normality, Heteroscedasticity, Autocorrelation and Multicollinearity. This, however, does not imply that the other assumptions of multiple regression analysis are of less importance.

Tests for the Assumptions of Regression Analysis

It is usually expected that the tests for the assumptions of regression analysis be conducted first before the regression analysis is carried out because it is the most important aspect of regression analysis which indicates that the model will be perfectly fitted.

a) Test for the Normality Assumption

One of the assumptions required by OLS method for the estimability of the parameters in the regression model is that the error terms are normally distributed. Gujarati [7] stated that a simple graphical representation (either a histogram of residuals or a normal probability plot) can be used to explain whether the residuals are normally distributed. With the histogram of residuals, the shape of the normal distribution curve can be ascertained on it; while for normal probability plot, the Anderson-Darling test will be used to study the shape of the probability density function of the random variables.

The procedure for the Anderson-Darling test for the normality assumption are very much discussed and outlined in, for example, [17] and [7].

b) Test for Homoscedasticity Assumption

Homoscedasticity or equal variance of the error term is another assumption required by the OLS method for the estimability of the parameters in the regression model. In order to confirm the existence of heteroscedasticity, some commonly used tests are namely; Breusch– Pagan (sometimes referred to as Breusch– Pagan– Godfrey) test, Spearman Rank Correlation test and Goldfeld-Quandt test.

The Spearman Rank Correlation test is simple and it is applicable to data with small and large sample sizes. The Goldfeld-Quandt test is applicable when the number of observations is greater than twice the number of independent variables. The success of the Goldfeld-Quandt test depends on the value of the middle observations being omitted and identifying the correct X-variable with which to order the observations. This limitation of the Goldfeld-Quandt test can be avoided if the Breusch– Pagan test is considered [7].

For the procedure of conducting the Breusch– Pagan– Godfrey test, the Spearman Rank Correlation test and the Goldfeld-Quandt test for the heteroscedasticity assumption, see, for example, [7] and [14].

c) Test for Multicollinearity Assumption

Some common tests for multicollinearity include the Farrar-Glauber test and Variance Inflation Factor (VIF); with the VIF being the most prominent in terms of usage. The VIF measures how much the variance of the estimated regression parameters are inflated as compared to when the independent variables are not linearly related (see, for example, Yoo et al, [19]). The null hypothesis of 'no perfect multicollinearity among the independent variables' is to be rejected if and only if calculated test statistic, VIF, is greater than or equal to 10 (see, for example, Hair et al, [8]; Rawlings et al, [16]).

The procedures of carrying out the three-stage Farrar-Glauber test and the VIF are as outlined in; see for example, [11], [8] and [16].

d) Test for Autocorrelation Assumption

The usually employed test for testing for the autocorrelation assumption in regression analysis is the Durbin-Watson test. The procedure of this test is outlined in; see for example, [11]. According to Koutsoyiannis [11], the null hypothesis for the Durbin-Watson test of 'no autocorrelation' is to be rejected if and only if the calculated Durbin-Watson test statistic, DW, is not approximately equal to 2.

Remediations to Unsatisfied Assumption(s) of Regression Analysis

e) Remedying the Incidence of Heteroscedasticity

The presence of heteroscedasticity in the multiple 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 [7] stated that this lack of efficiency makes the outcome of the usual hypothesis-testing to be dubious. Gujarati [7] further elaborated that there are two approaches to remediation; the first one is when the error variance, σ_i^2, is known and the second one is when σ_i^2 is unknown. According to Gujarati [7], the Weighted Least Square (WLS) approach, the Feasible Generalized Least Square (FGLS) approach, etc., can be applied to remedy the heteroscedasticity in

the regression model if the error variance, σ_i^2, is known; while Log-transformation of the variables, Inverse and Square root transformations, etc., can be applied for correcting the heteroscedasticity in the regression model if the error variance, σ_i^2, is unknown.

However, in this study, the Log-transformation method of correcting the presence of heteroscedasticity shall be employed where and when the need arises. The procedure of the Log-transformation method involves taking the natural logarithmic of each entry in the dataset and then applying the Ordinary Least Square (OLS) method (see, Gujarati [7]). According to Gujarati [7], a log-transformation such as,

$$InY_i = \beta_1 + \beta_2 InX_i + e_i \tag{2.8}$$

very often reduces heteroscedasticity when compared with the regression,

$$Y_i = \beta_1 + \beta_2 X_i + e_i \tag{2.9}$$

f) Remedying the Incidence of Multicollinearity

Koutsoyiannis [11] stated that the serious effects of the existence of multicollinearity on the estimates of the regression model coefficients may be remedied by adopting any of the followings;

i) Application of Method Incorporating Extraneous Quantitative Information;
ii) Increase of the Size of the Sample;
iii) Substitution of Lagged Variables for other Explanatory Variables in Distributed-Lag Models;
iv) Introduction of Additional Equation in the Model; or
v) Application of the Principal Component Method.

It may be interesting to know that if the regression model parameters estimation is mainly for forecasting purposes, the incidence and the consequent remediation of multicollinearity in the data may be ignored. According to Koutsoyiannis [11], the estimates of the original model may be accepted despite the existence of multicollinearity, only if the purpose of the estimation is for forecast, and provided that the same pattern of multicollinearity of the independent variables continue in the period of prediction. Koutsoyiannis [11] further added that, in such a case, if one tries to remove the independent variables responsible for multicollinearity, it will lead to specification bias.

3. Data and analyses

This study showcased two different HYPOTHETICAL datasets (Data A and B) just for the purpose of illustrating the very context of this research. Data A (the Original), having six (6) predictor variables and one response variable, are as presented in Columns 1 to 7 of Table 3.1 (see Appendix A); while Data B (the Original), having five (5) predictor variables and one response variable, are as presented in Columns 1 to 6 of Table 3.2 (see Appendix B).

Adopting (2.1), this study uses the following theoretical model to assess the independent variables that are associated with the dependent variable; and for Data A and B, the multiple regression equations will, respectively, be given as,

$$Y_A = \beta_{0A} + \beta_{1A} X_{1A} + \beta_{2A} X_{2A} + \beta_{3A} X_{3A} + \beta_{4A} X_{4A} + \beta_{5A} X_{5A} + \beta_{6A} X_{6A} + e_A \tag{2.10}$$

And

$$Y_B = \beta_{0B} + \beta_{1B} X_{1B} + \beta_{2B} X_{2B} + \beta_{3B} X_{3B} + \beta_{4B} X_{4B} + \beta_{5B} X_{5B} + e_B \tag{2.11}$$

The data analyses in this study shall be done with the aid of the following statistical packages; Microsoft Office Excel (2016), Minitab (2019), SPSS version 26, and NCSS (2012). The results outputs from the various computer packages employed in testing the relevant assumptions of the multiple linear regression and correlation analyses, as well as the main data analyses are as presented in Sub-sections 3.1 and 3.2.

3.1. Analyses on data a

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 A is hereby presented in this sub-section.

3.1.1. Analyses on data a (the original)

The procedure of the Multiple Linear Regression Analysis (MLRA) is carried out on Data A (the Original). The results outputs of each stage of this procedure are presented in Tables 3.3 to 3.9 and Figure 3.1.

3.1.1.1. Descriptive statistics for data a (the original)

The descriptive statistics for Data A (the Original), which include the mean, the standard deviation, and the minimum and the maximum values for each of the six independent variables and the dependent variable are presented in Table 3.3.

Table 3.3: Descriptive Statistics for Data A (the Original)

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
X_{1A}	60	12366.34	16251.95	71.02	58472.88
X_{2A}	60	-85.42717	762.2006	-5889.73	99.6
X_{3A}	60	13.93617	6.263899	6	29.8
X_{4A}	60	3506.791	6128.483	13.52	28729.56

X_{5A}	60	9.136666	6.539916	1.9	26.4
X_{6A}	60	15.985	5.255791	8.46	30.4
Y_A	60	66.2605	93.05166	0.55	358.81

3.1.1.2. Testing for the normality assumption on data a (the original)

Data A (the Original) were tested for the normality assumption using the Anderson-Darling test, Shapiro-Wilk test and the d' Agostino-Pearson test. The results outputs of these tests are presented in Figure 3.1 and Table 3.4, respectively.

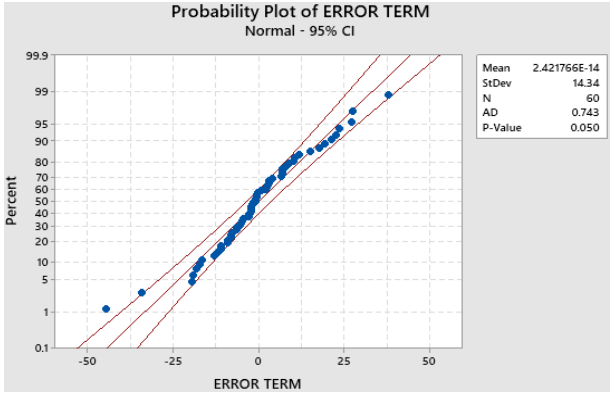

Fig. 3.1: The Anderson-Darling Test for the Normality Assumption on Data A (The Original).

Table 3.4: Shapiro-Wilk Test and d Agostino-Pearson Test for Normality Assumption on Data A (the Original)

Shapiro-Wilk Test		D Agostino-Pearson Test	
W-stat	0.967586	DA-stat	3.828925
P-value	0.111069	P-value	0.147421
alpha	0.05	alpha	0.05
Normal	yes	Normal	yes

3.1.1.3. Testing for the heteroscedasticity assumption on data a (the original)

Data A (the Original) were tested for the heteroscedasticity assumption using Breusch-Pagan test and White test for heteroscedasticity. The results outputs of this test are presented in Table 3.5.

Table 3.5: Breusch-Pagan Test and White Test for Heteroscesdasticity Assumption on Data A (the Original)

	Breusch-Pagan Test	White Test
LM stat	20.26848	18.43241
df	6	2
P-value	0.00248	9.93E-05
F stat	4.506202	12.63782
df1	6	2
df2	53	57
P-value	0.000929	2.87E-05

3.1.1.4. Multiple linear regression analysis for data a (the original)

The results outputs for the multiple linear regression analysis on Data A (the Original) are presented in Tables 3.6 to 3.8.

Table 3.6: Regression Model Coefficients for Data A (the Original)

Variable	Unstandardized Coefficients		Standardized Coefficients	t-stat	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
Intercept	-8.537	10.363		-.824	.414					
X_{1A}	.001	.000	.226	6.341	.000	.797	.657	.134	.353	2.837
X_{2A}	-.001	.003	-.004	-.203	.840	.092	-.028	-.004	.974	1.026
X_{3A}	.971	.392	.065	2.475	.017	.398	.322	.052	.642	1.557
X_{4A}	.010	.000	.674	23.355	.000	.941	.955	.494	.539	1.857
X_{5A}	2.827	.484	.199	5.842	.000	.782	.626	.124	.387	2.581
X_{6A}	-1.032	.494	-.058	-2.092	.041	.328	-.276	-.044	.577	1.734

Table 3.7: Model Summary for Data A (the Original)

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	
.988	.976	.974	15.133	.976	362.969	6	53	.000	1.931

Table 3.8: Additional Overall Fit of the Regression Model for Data A (the Original)

	AIC	332.5929
	AICc	335.4165
	SBC	347.2533

3.1.1.5. Multiple linear correlation analysis for data a (the original)

The results outputs for the multiple linear correlation analysis on Data A (the Original) are presented in Table 3.9.

Table 3.9: Correlations for Data A (the Original)

	Variable	Y_A	X_{1A}	X_{2A}	X_{3A}	X_{4A}	X_{5A}	X_{6A}
Correlations	Y_A	1.000	0.797	0.092	0.398	0.941	0.782	0.328
	X_{1A}	0.797	1.000	0.099	0.363	0.632	0.738	0.440
	X_{2A}	0.092	0.099	1.000	0.153	0.072	0.076	-0.014
	X_{3A}	0.398	0.363	0.153	1.000	0.272	0.269	-0.248
	X_{4A}	0.941	0.632	0.072	0.272	1.000	0.631	0.315
	X_{5A}	0.782	0.738	0.076	0.269	0.631	1.000	0.457
	X_{6A}	0.328	0.440	-0.014	-0.248	0.315	0.457	1.000

3.1.2. Analyses on data a (now with heteroscedasticity remedied)

The procedure of the Multiple Linear Regression Analysis is carried out on Data A which failed the heteroscedasticity assumption but is now corrected. The results outputs of the procedure are presented in Tables 3.10 to 3.16 and Figure 3.2.

3.1.2.1. Descriptive statistics for data a (now with heteroscedasticity remedied)

The descriptive statistics for Data A (Now with Heteroscedasticity Remedied), which include the mean, the standard deviation, and the minimum and the maximum values for each of the six independent variables and the dependent variable are presented in Table 3.10.

Table 3.10: Descriptive Statistics for Data A (Now with Heteroscedasticity Remedied)

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
InX_{1A}	60	8.082163	2.032619	4.262961	10.97632
InX_{2A}	60	2.19597	1.017399	-0.1508229	4.601162
InX_{3A}	60	2.525065	0.4845554	1.791759	3.394508
InX_{4A}	60	6.546892	2.001891	2.60417	10.26568
InX_{5A}	60	1.979515	0.6837802	0.6418539	3.273364
InX_{6A}	60	2.720674	0.3196657	2.135849	3.415223
InY_A	60	2.22636	2.463846	-0.597837	5.882793

3.1.2.2. Testing for the normality assumption on data a (now with heteroscedasticity remedied)

Data A (Now with Heteroscedasticity Remedied) were tested for the normality assumption using the Anderson-Darling test, Shapiro-Wilk test and the d' Agostino-Pearson test. The results outputs of these tests are presented in Figure 3.2 and Table 4.11, respectively.

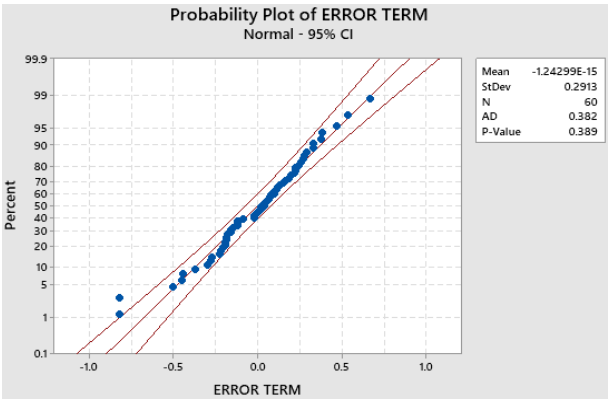

Fig. 3.2: The Anderson-Darling Test for the Normality Assumption on Data A (Now with Heteroscedasticity Remedied)

Table 3.11: Shapiro-Wilk Test and d Agostino-Pearson Test for Normality Assumption on Data A (Now with Heteroscedasticity Remedied)

Shapiro-Wilk Test		d Agostino-Pearson Test	
W-stat	0.975292	DA-stat	4.584655
P-value	0.262465	P-value	0.101031
alpha	0.05	alpha	0.05
Normal	yes	Normal	yes

3.1.2.3. Testing for the heteroscedasticity assumption on data a (now with heteroscedasticity remedied)

Data A (Now with Heteroscedasticity Remedied) were tested for the heteroscedasticity assumption using Breusch-Pagan test and White test for heteroscedasticity. The results outputs of this test are presented in Table 3.12.

Table 3.12: Breusch-Pagan Test and White Test for Heteroscesdasticity Assumption on Data A (Now with Heteroscedasticity Remedied)

	Breusch-Pagan Test	White Test
LM stat	7.08136	2.823156
df	6	2
P-value	0.313389	0.243758
F stat	1.182041	1.407212
df1	6	2
df2	53	57
P-value	0.330032	0.2532

3.1.2.4. Multiple linear regression analysis for data a (now with heteroscedasticity remedied)

The results outputs for the multiple linear regression analysis on Data A (Now with Heteroscedasticity Remedied) are presented in Tables 3.13 to 3.15.

Table 3.13: Regression Model Coefficients for Data A (Now with Heteroscedasticity Remedied)

Variable	Unstandardized Coefficients		Standardized Coefficients	t-stat	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
Intercept	-5.658	.574		-9.852	.000					
InX_{1A}	.319	.045	.263	7.027	.000	.852	.694	.114	.188	5.320
InX_{2A}	-.234	.055	-.097	-4.263	.000	.333	-.505	-.069	.511	1.956
InX_{3A}	1.226	.161	.241	7.621	.000	.811	.723	.124	.264	3.794
InX_{4A}	.707	.040	.574	17.680	.000	.945	.925	.287	.250	4.000
InX_{5A}	.384	.091	.107	4.218	.000	.614	.501	.069	.413	2.419
InX_{6A}	-.978	.176	-.127	-5.558	.000	.077	-.607	-.090	.506	1.977

Table 3.14: Model Summary for Data A (Now with Heteroscedasticity Remedied)

Multiple R	R Square	Adjusted R Square	Std. Error of the Esti-mate	Change Statistics					Durbin -Watson
				R Square Change	F Change	df1	df2	Sig. F Change	
.993	.986	.984	.307	.986	623.148	6	53	.000	1.791

Table 3.15: Additional Overall Fit of the Regression Model for Data A (Now with Heteroscedasticity Remedied)

AIC	-135.021
AICc	-132.198
SBC	-120.361

3.1.2.5. Multiple linear correlation analysis for data a (now with heteroscedasticity remedied)

The results outputs for the multiple linear correlation analysis on Data A (Now with Heteroscedasticity Remedied) are presented in Table 3.16.

Table 3.16: Correlations for Data A (Now with Heteroscedasticity Remedied)

	Variable	InY_A	InX_{1A}	InX_{2A}	InX_{3A}	InX_{4A}	InX_{5A}	InX_{6A}
Correlations	InY_A	1.000	.852	.333	.811	.945	.614	.077
	InX_{1A}	.852	1.000	.577	.692	.763	.616	.206
	InX_{2A}	.333	.577	1.000	.470	.228	.166	-.123
	InX_{3A}	.811	.692	.470	1.000	.652	.241	-.262
	InX_{4A}	.945	.763	.228	.652	1.000	.637	.258
	InX_{5A}	.614	.616	.166	.241	.637	1.000	.495
	InX_{6A}	.077	.206	-.123	-.262	.258	.495	1.000

3.2. Analyses on data b

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 B is hereby presented in this sub-section.

3.2.1. Analyses on data b (the original)

The procedure of the Multiple Linear Regression Analysis is carried out on Data B (the Original). The results outputs of the procedure are presented in Tables 3.17 to 3.23 and Figure 3.3.

3.2.1.1. Descriptive statistics for data b (the original)

The descriptive statistics for Data B (the Original), which include the mean, the standard deviation, and the minimum and the maximum values for each of the six independent variables and the dependent variable are presented in Table 3.17

Table 3.17: Descriptive Statistics for Data B (the Original)

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
X_{1B}	57	7.087719	4.128649	2	16
X_{2B}	57	9.631579	6.744393	2	30
X_{3B}	57	4.456141	2.315092	2	12
X_{4B}	57	4.456141	2.57786	2	12
X_{5B}	57	58.42105	31.51187	5	100
Y_B	57	2.407526	2.506772	0.201	9.21

3.2.1.2. Testing for the Normality Assumption on Data B (the Original)

Data B (the Original) were tested for the normality assumption using the Anderson-Darling test, Shapiro-Wilk test and the d' Agostino-Pearson test. The results outputs of these tests are presented in Figure 4.3 and Table 3.18, respectively.

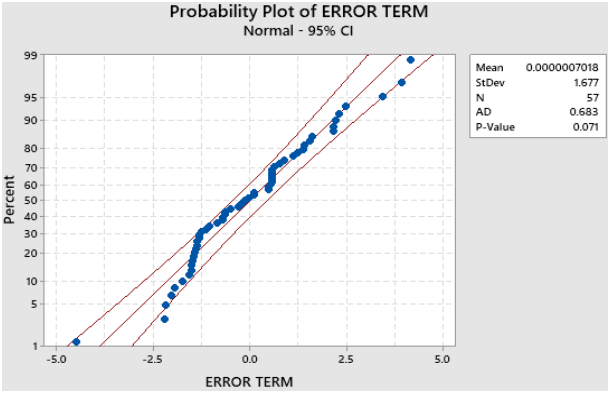

Fig. 3.3: The Anderson-Darling Test for the Normality Assumption on Data B (the Original).

Table 3.18: Shapiro-Wilk Test and d Agostino-Pearson Test for Normality Assumption on Data B (the Original)

Shapiro-Wilk Test		D Agostino-Pearson Test	
W-stat	0.966699	DA-stat	1.76228
P-value	0.117574	P-value	0.41431
alpha	0.05	alpha	0.05
Normal	yes	Normal	yes

3.2.1.3. Testing for the heteroscedasticity assumption on data b (the original)

Data B (the Original) were tested for the heteroscedasticity assumption using Breusch-Pagan test and White test for heteroscedasticity. The results outputs of this test are presented in Table 3.19.

Table 3.19: Breusch-Pagan Test and White Test for Heteroscesdasticity Assumption on Data B (the Original)

	Breusch-Pagan Test	White Test
LM stat	14.43118	7.522686
df	5	2
P-value	0.01309	0.023252
F stat	3.457885	12.63782
df1	5	2
df2	51	54
P-value	0.009114	0.021895

3.2.1.4. Multiple linear regression analysis for data b (the original)

The results outputs for the multiple linear regression analysis on Data B (the Original) are presented in Tables 3.20 to 3.22.

Table 3.20: Regression Model Coefficients for Data B (the Original)

Variable	Unstandardized Coefficients		Standardized Coefficients	t-stat	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
Intercept	3.979	.989		4.023	.000					
X_{1B}	-.274	.061	-.451	-4.489	.000	-.326	-.532	-.420	.868	1.152
X_{2B}	.155	.037	.417	4.232	.000	.463	.510	.396	.905	1.105
X_{3B}	.279	.116	.258	2.400	.020	.300	.319	.225	.761	1.314
X_{4B}	-.482	.102	-.496	-4.724	.000	-.262	-.552	-.442	.796	1.256
X_{5B}	-.004	.008	-.046	-.490	.626	-.118	-.068	-.046	.975	1.026

Table 3.21: Model Summary for Data B (the Original)

Multiple R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics R Square Change	F Change	df1	df2	Sig. F Change	Durbin -Watson
.743	.553	.509	1.757	.553	12.604	5	51	.000	1.710

Table 4.22: Additional Overall Fit of the Regression Model for Data B (the Original)

AIC	69.89669
AICc	72.1824
SBC	82.15499

3.2.1.5. Multiple linear correlation analysis for data a (the original)

The results outputs for the multiple linear correlation analysis on Data B (the Original) are presented in Table 3.23.

Table 3.23: Correlations for Data B (the Original)

	Variable	Y_B	X_{1B}	X_{2B}	X_{3B}	X_{4B}	X_{5B}
Correlations	Y_B	1.000	-.326	.463	.300	-.262	-.118
	X_{1B}	-.326	1.000	.063	-.240	-.321	-.037
	X_{2B}	.463	.063	1.000	.266	-.006	-.073
	X_{3B}	.300	-.240	.266	1.000	.365	-.105
	X_{4B}	-.262	-.321	-.006	.365	1.000	.062
	X_{5B}	-.118	-.037	-.073	-.105	.062	1.000

3.2.2. Analyses on data b (now with heteroscedasticity remedied)

The procedure of the Multiple Linear Regression Analysis is carried out on Data B which failed the heteroscedasticity assumption but is now corrected. The results outputs of the procedure are presented in Tables 3.24 to 3.30 and Figure 3.4.

3.2.2.1. Descriptive statistics for data b (now with heteroscedasticity remedied)

The descriptive statistics for Data B (Now with Heteroscedasticity Remedied), which include the mean, the standard deviation, and the minimum and the maximum values for each of the five independent variables and the dependent variable are presented in Table 3.24.

Table 3.24: Descriptive Statistics for Data B (Now with Heteroscedasticity Remedied)

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
InX_{1B}	57	1.77396	0.6380438	0.6931472	2.772589
InX_{2B}	57	2.05582	0.6520647	0.6931472	3.401197
InX_{3B}	57	1.36708	0.5081612	0.6931472	2.484907
InX_{4B}	57	1.360702	0.5008541	0.6931472	2.484907
InX_{5B}	57	3.805715	0.8916203	1.609438	4.60517
InY_B	57	0.3286631	1.116958	-1.60445	2.22029

3.2.2.2. Testing for the normality assumption on data b (now with heteroscedasticity remedied)

Data B (Now with Heteroscedasticity Remedied) were tested for the normality assumption using the Anderson-Darling test, Shapiro-Wilk test and the d' Agostino-Pearson test. The results outputs of these tests are presented in Figure 3.4 and Table 3.25, respectively.

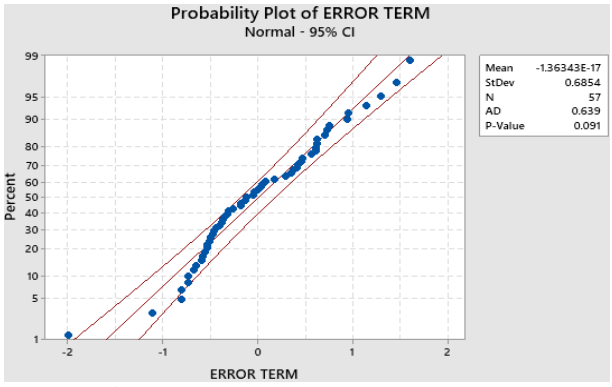

Fig. 4.4: The Anderson-Darling Test for the Normality Assumption on Data B (Now with Heteroscedasticity Remedied).

Table 3.25: Shapiro-Wilk Test and d Agostino-Pearson Test for Normality Assumption on Data B (Now with Heteroscedasticity Remedied)

Shapiro-Wilk Test		D Agostino-Pearson Test	
W-stat	0.970997	DA-stat	0.772837
P-value	0.186676	P-value	0.679486
alpha	0.05	alpha	0.05
Normal	yes	Normal	yes

3.2.2.3. Testing for heteroscedasticity assumption on data b (now with heteroscedasticity remedied)

Data B (Now with Heteroscedasticity Remedied) were tested for the heteroscedasticity assumption using Breusch-Pagan test and White test for heteroscedasticity. The results outputs of this test are presented in Table 3.26.

Table 3.26: Breusch-Pagan Test and White Test for Heteroscesdasticity Assumption on Data B (Now with Heteroscedasticity Remedied)

	Breusch-Pagan Test	White Test
LM stat	8.690418	0.560865
df	5	2
P-value	0.122068	0.755457
F stat	1.83488	0.268313
df1	5	2
df2	51	54
P-value	0.122598	0.765682

3.2.2.4. Multiple linear regression analysis for data b (now with heteroscedasticity remedied)

The results outputs for the multiple linear regression analysis on Data B (Now with Heteroscedasticity Remedied) are presented in Tables 3.27 to 3.29.

Table 3.27: Regression Model Coefficients for Data B (Now with Heteroscedasticity Remedied)

Variable	Unstandardized Coefficients		Standardized Coefficients	t-stat	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
Intercept	.093	.703		.132	.896					
InX_{1B}	-.844	.166	-.482	-5.080	.000	-.336	-.580	-.436	.820	1.220
InX_{2B}	.917	.169	.535	5.411	.000	.512	.604	.465	.754	1.326
InX_{3B}	.620	.255	.282	2.431	.019	.345	.322	.209	.548	1.824
InX_{4B}	-1.031	.230	-.462	-4.479	.000	-.185	-.531	-.385	.692	1.444
InX_{5B}	.106	.110	.085	.967	.338	.051	.134	.083	.962	1.039

Table 3.28: Model Summary for Data B (Now with Heteroscedasticity Remedied)

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	
.790	.624	.587	.718	.624	16.892	5	51	.000	1.781

Table 3.29: Additional Overall Fit of the Regression Model for Data B (Now with Heteroscedasticity Remedied)

AIC	-32.0801
AICc	-29.7944
SBC	-19.8218

3.2.2.5. Multiple linear correlation analysis for data b (now with heteroscedasticity remedied)

The results outputs for the multiple linear correlation analysis on Data B (Now with Heteroscedasticity Remedied) are presented in Table 3.30.

Table 3.30: Correlations for Data B (With Heteroscedasticity Remedied)

	Variable	InY_B	InX_{1B}	InX_{2B}	InX_{3B}	InX_{4B}	InX_{5B}
Correlations	InY_B	1.000	-.336	.512	.345	-.185	.051
	InX_{1B}	-.336	1.000	.203	-.268	-.242	.014
	InX_{2B}	.512	.203	1.000	.361	.045	-.071
	InX_{3B}	.345	-.268	.361	1.000	.527	-.184
	InX_{4B}	-.185	-.242	.045	.527	1.000	-.136
	InX_{5B}	.051	.014	-.071	-.184	-.136	1.000

4. Discussions on results from data analyses

The results obtained in the multiple linear regression and correlation analyses of the hypothetical Data A and B are put up for discussions in this section.

4.1. Discussions on the results of the analyses on data a (the original)

Some of the descriptive statistics for Data A (the Original) are presented in Table 3.3, which include the count, mean, standard deviation, minimum and maximum values for each of the variables that are to be involved in the multiple linear regression and correlation analyses. The normality and heteroscedasticity assumptions were tested on Data A (the Original) prior to the conduction of the multiple linear regression and correlation analyses.

The p-Values of the Anderson-Darling test (in Figure 3.1) and the Shapiro-Wilk and d' Agostino-Pearson tests (both in Table 3.4) which are 0.050, 0.111069 and 0.147421, respectively, are all indicative that Data A (the Original) satisfied the normality assumption. Both the

Breusch-Pagan test and the White test (in Table 3.5) yielded p-Values of 0.00248 and 9.93E-0.5, respectively, which imply that Data A (the Original) failed the heteroscedasticity assumption.
Despite Data A (the Original) failing the heteroscedasticity assumption, the multiple linear regression and correlation analyses were still carried out on the hypothetical Data A (the Original). The results outputs of the multiple linear regression and correlation analyses on Data A (the Original) are presented in Tables 3.6 to 3.9.
From Table 3.6, the multiple linear regression model for Data A (the Original) is obtained as,

$$\hat{Y}_A = -8.537 + 0.001X_{1A} - 0.001X_{2A} + 0.971X_{3A} + 0.010X_{4A} + 2.827X_{5A} - 1.032X_{6A} \quad (4.1)$$

It was evident, also, from Table 3.6 that all the independent variables except X_{2A} are significant as their p-Values are all less than the chosen level of significance, α =0.05. Also shown in Table 3.6 are the values of the Variance Inflation Factor (VIF) for all the six independent variables are all less than the critical value, 10; which is an indication that the multicollinearity assumption was satisfied in Data A (the Original).
From Table 3.7, the value of the computed Durban-Watson statistic of 1.931 (which is approximately equal to 2) implies that the autocorrelation assumption was satisfied in Data A (the Original). Also, from Table 3.7, the computed F-statistic of 362.969 (a p-Value equivalent of about 0.000) led to the conclusion that the model is of good-fit to Data A (the Original); and the R-square value of 0.976 showed that about 97.6% of the total variation in the dependent variable, Y_A, is being accounted for by the variations in the independent variables, X_{1A}, X_{2A}, X_{3A}, X_{4A}, X_{5A} and X_{6A}; while about 2.4% is left unaccounted for perhaps by some other variables not included in the modelling.
The values of the R-square (=0.976) and Adjusted R-square (=0.974) in Table 3.7 indicate the level of adequacy of the established regression model for Data A (the Original). Also, the values of the AIC, AICc and SBC (=332.5929, 335.4165 and 347.2533, respectively) in Table 3.8 are additional indication of the overall fit for the same established regression model for Data A (the Original).
The value of the Multiple R (=0.988) in the multiple linear regression analysis on Data A (the Original), presented in Table 3.7, showed that there was a strong positive correlation among the dependent variable, Y_A, and the independent variables, X_{1A}, X_{2A}, X_{3A}, X_{4A}, X_{5A} and X_{6A}. From the results outputs of the multiple linear correlation analysis on Data A (the Original), presented in Table 3.9, it was evident that the dependent variable, Y_A, specifically has strong positive correlations with X_{4A}, X_{1A} and X_{5A} (in that order of magnitude); weak positive correlations with X_{3A} and X_{6A} (in that order of magnitude); and no correlation at all with X_{2A}.
Finally, the predicted values of the dependent variable obtained using (4.1) are presented in Column 8 of Table 3.1 (see Appendix A). Also, Figure (4.1) presents the graph of these predicted values superimposed with the graph of the original values of the dependent variables (see Appendix C).

4.2. Discussions on the results of the analyses on data a (now with heteroscedasticity remedied)

The failure of the heteroscedasticity assumption in Data A (the Original) necessitated the correction or the remediation of the data for the presence of heteroscedasticity. The correction is done by employing the Log-transformation method (see Gujarati [7]) as expressed in (2.8); which in this case is given by,

$$InY_A = \beta_{0A} + \beta_{1A}InX_{1A} + \beta_{2A}InX_{2A} + \beta_{3A}InX_{3A} + \beta_{4A}InX_{4A} + \beta_{5A}InX_{5A} + \beta_{6A}InX_{6A} + e_A \quad (4.2)$$

Going forward with the usual procedure, Table 3.10 presents some of the descriptive statistics for Data A (Now with Heteroscedasticity Remedied). The normality and heteroscedasticity assumptions were tested on Data A (Now with Heteroscedasticity Remedied) prior to the conduction of the multiple linear regression and correlation analyses.
The p-Values of the Anderson-Darling test (in Figure 3.2) and the Shapiro-Wilk and d' Agostino-Pearson tests (both in Table 3.11) which are 0.389, 0.262465 and 0.101031, respectively, are all indicative that Data A (Now with Heteroscedasticity Remedied) satisfied the normality assumption. Both the Breusch-Pagan test and the White test (in Table 3.12) yielded p-Values of 0.313389 and 0.243758, respectively, which imply that Data A (Now with Heteroscedasticity Remedied) now satisfy the heteroscedasticity assumption.
The multiple linear regression and correlation analyses are now carried out on Data A (Now with Heteroscedasticity Remedied), and the results outputs are presented in Tables 3.13 to 3.16. From Table 3.13, the multiple linear regression model for Data A (Now with Heteroscedasticity Remedied) is obtained as,

$$In\hat{Y}_A = -5.658 + 0.319InX_{1A} - 0.234InX_{2A} + 1.226InX_{3A} + 0.707InX_{4A} + 0.384InX_{5A} - 0.978InX_{6A} \quad (4.3)$$

It was evident, also, from Table 3.13 that all the independent variables are significant as their p-Values are all less than the chosen level of significance, α =0.05. Also shown in Table 3.13 are the values of the Variance Inflation Factor (VIF) for all the six independent variables are all less than the critical value, 10; which is an indication that the multicollinearity assumption was satisfied in Data A (Now with Heteroscedasticity Remedied).
From Table 3.14, the value of the computed Durban-Watson statistic of 1.791 (which is approximately equal to 2) implies that the autocorrelation assumption was satisfied in Data A (Now with Heteroscedasticity Remedied). Also, the computed F-statistic of 623.148 (a p-Value equivalent of about 0.000) led to the conclusion that the model is of good-fit to Data A (Now with Heteroscedasticity Remedied); and the R-square value of 0.986 showed that about 98.6% of the total variation in the dependent variable, InY_A, is being accounted for by the variations in the independent variables, InX_{1A}, InX_{2A}, InX_{3A}, InX_{4A}, InX_{5A} and InX_{6A}; while about 1.4% is left unaccounted for perhaps by some other variables not included in the modelling.
The values of the R-square (=0.986) and Adjusted R-square (=0.984) in Table 3.14 indicate the level of adequacy of the established regression model for Data A (Now with Heteroscedasticity Remedied). Also, the values of the AIC, AICc and SBC (= -135.021, -132.198 and -

120.361, respectively) in Table 3.15 are additional indication of the overall fit for the same established regression model for Data A (Now with Heteroscedasticity Remedied).

The value of the Multiple R (=0.993) in the multiple linear regression analysis on Data A (Now with Heteroscedasticity Remedied), presented in Table 3.14, showed that there was a very strong positive correlation among the dependent variable, $In\mathrm{Y}_A$, and the independent variables, InX_{1A}, InX_{2A}, InX_{3A}, InX_{4A}, InX_{5A} and InX_{6A}. From the results outputs of the multiple linear correlation analysis on Data A (Now with Heteroscedasticity Remedied), presented in Table 3.16, it was evident that the dependent variable, $In\mathrm{Y}_A$, specifically has strong positive correlations with InX_{4A}, InX_{1A} and InX_{3A} (in that order of magnitude); average positive correlation with InX_{5A}; weak positive correlation with InX_{2A}; and no correlation at all with InX_{6A}.

Finally, the predicted values of the dependent variable obtained using (4.3) as well as their reversed-transformed values are, respectively, presented in Columns 9 and 10 of Table 3.1 (see Appendix A). Also, Figure (4.2) presents the graph of these reversed-transformed (supposed real) predicted values superimposed with the graph of the original values of the dependent variables (see Appendix C).

4.3. Discussions on the results of the analyses on data b (the original)

The count, mean, standard deviation, minimum and maximum values of each of the variables in Data B (the Original) that are to be included in the multiple linear regression and correlation analyses are presented in Table 3.17. Also, the normality and heteroscedasticity assumptions were tested on Data A (Now with Heteroscedasticity Remedied) prior to the conduction of the multiple linear regression and correlation analyses. The p-Values of the Anderson-Darling test (in Figure 3.3) and the Shapiro-Wilk and d' Agostino-Pearson tests (both in Table 3.18) which are 0.071, 0.117574 and 0.41431, respectively, are all indicative that Data B (the Original) satisfied the normality assumption. Both the Breusch-Pagan test and the White test (in Table 3.19) yielded p-Values of 0.01309 and 0.023252, respectively, which imply that Data B (the Original) failed the heteroscedasticity assumption.

Although the test for the heteroscedasticity assumption failed, the multiple linear regression and correlation analyses were still carried out on the hypothetical Data B (the Original). The results outputs of the multiple linear regression and correlation analyses on Data B (the Original) are presented in Tables 3.20 to 3.23. From Table 3.20, the multiple linear regression model for Data B (the Original) is obtained as,

$$\hat{\mathrm{Y}}_B = 3.979 - 0.274X_{1B} + 0.155X_{2B} + 0.279X_{3B} - 0.482X_{4B} - 0.004X_{5B} \tag{4.4}$$

It was evident, also, from Table 3.20 that all the independent variables except X_{5B} are significant as their p-Values are all less than the chosen level of significance, α =0.05. Also shown in Table 3.20 are the values of the Variance Inflation Factor (VIF) for all the five independent variables are all less than the critical value, 10; which is an indication that the multicollinearity assumption was satisfied in Data B (the Original).

From Table 3.21, the value of the computed Durban-Watson statistic of 1.710 (which is approximately equal to 2) implies that the autocorrelation assumption was satisfied in Data B (the Original). Also, from Table 3.21, the computed F-statistic of 12.604 (a p-Value equivalent of about 0.000) led to the conclusion that the model is of good-fit to Data B (the Original); and the R-square value of 0.553 showed that about 55.3% of the total variation in the dependent variable, Y_B, is being accounted for by the variations in the independent variables, X_{1B}, X_{2B}, X_{3B}, X_{4B} and X_{5B}; while about 22.7% is left unaccounted for perhaps by some other variables not included in the modelling.

The values of the R-square (=0.553) and Adjusted R-square (=0.509) in Table 3.21 indicate the level of adequacy of the established regression model for Data B (the Original). Also, the values of the AIC, AICc and SBC (=69.89669, 72.1824 and 82.15499, respectively) in Table 3.22 are additional indication of the overall fit for the same established regression model for Data B (the Original).

The value of the Multiple R (=0.743) in the multiple linear regression analysis on Data A (Now with Heteroscedasticity Remedied), presented in Table 3.21, showed that there was a positive correlation among the dependent variable, Y_B, and the independent variables, X_{1B}, X_{2B}, X_{3B}, X_{4B} and X_{5B}. From the results outputs of the multiple linear correlation analysis on Data B (the Original), presented in Table 3.23, it was evident that the dependent variable, Y_B, specifically has an average positive correlation with X_{2B}; weak positive correlation with X_{3B}; and weak negative correlations with X_{1B}, X_{4B} and X_{5B} (in that order of magnitude).

Finally, the predicted values of the dependent variable obtained using (4.4) are presented in Column 7 of Table 3.2 (see Appendix B). Also, Figure (4.3) presents the graph of these predicted values superimposed with the graph of the original values of the dependent variables (see Appendix C).

4.4. Discussions on the results of the analyses on data b (now with heteroscedasticity remedied)

The failure of the heteroscedasticity assumption in Data B (the Original) necessitated the correction or the remediation of the data for the presence of heteroscedasticity. The correction is done by employing the Log-transformation method (see Gujarati [7]) as expressed in (2.8); which in this case is given by,

$$In\mathrm{Y}_B = \beta_{0B} + \beta_{1B}InX_{1B} + \beta_{2B}InX_{2B} + \beta_{3B}InX_{3B} + \beta_{4B}InX_{4B} + \beta_{5B}InX_{5B} + e_B \tag{4.5}$$

Going forward with the usual procedure, Table 3.24 presents some of the descriptive statistics for Data B (Now with Heteroscedasticity Remedied). The normality and heteroscedasticity assumptions were tested on Data B (Now with Heteroscedasticity Remedied) prior to the conduction of the multiple linear regression and correlation analyses. The p-Values of the Anderson-Darling test (in Figure 3.4) and the Shapiro-Wilk and d' Agostino-Pearson tests (both in Table 3.25) which are 0.091, 0.186676 and 0.679486, respectively, are all indicative that Data B (Now with Heteroscedasticity Remedied) satisfied the normality assumption. Both the Breusch-Pagan test and the White test (in Table 3.26) yielded p-Values of 0.122068 and 0.755457, respectively, which imply that Data B (Now with Heteroscedasticity Remedied) now satisfy the heteroscedasticity assumption.

The multiple linear regression and correlation analyses are now carried out on Data B (Now with Heteroscedasticity Remedied), and the results outputs are presented in Tables 3.27 to 3.30. From Table 3.27, the multiple linear regression model for Data B (Now with Heteroscedasticity Remedied) is obtained as,

$$In\hat{Y}_B = 0.093 - 0.844InX_{1B} + 0.917InX_{2B} + 0.620InX_{3B} - 1.031InX_{4B} + 0.106InX_{5B} \quad (4.6)$$

It was evident, also, from Table 3.27 that all the independent variables except InX_{5B} are significant as their p-Values are all less than the chosen level of significance, α =0.05. Also shown in Table 3.27 are the values of the Variance Inflation Factor (VIF) for all the five independent variables are all less than the critical value, 10; which is an indication that the multicollinearity assumption was satisfied in Data B (Now with Heteroscedasticity Remedied).

From Table 3.28, the value of the computed Durban-Watson statistic of 1.781 (which is approximately equal to 2) implies that the autocorrelation assumption was satisfied in Data B (Now with Heteroscedasticity Remedied). Also, from Table 3.28, the computed F-statistic of 16.892 (a p-Value equivalent of about 0.000) led to the conclusion that the model is of good-fit to Data B (Now with Heteroscedasticity Remedied); and the R-square value of 0.624 showed that about 62.4% of the total variation in the dependent variable, $In\mathrm{Y}_B$, is being accounted for by the variations in the independent variables, InX_{1B}, InX_{2B}, InX_{3B}, InX_{4B} and InX_{5B}; while about 37.6% is left unaccounted for perhaps by some other variables not included in the modelling.

The values of the R-square (=0.624) and Adjusted R-square (=0.587) in Table 3.28 indicate the level of adequacy of the established regression model for Data B (Now with Heteroscedasticity Remedied). Also, the values of the AIC, AICc and SBC (= -32.0801, =29.7944 and -19.8218, respectively) in Table 3.29 are additional indication of the overall fit for the same established regression model for Data B (Now with Heteroscedasticity Remedied).

The value of the Multiple R (=0.790) in the multiple linear regression analysis on Data B (Now with Heteroscedasticity Remedied), presented in Table 4.28, showed that there was a strong positive correlation among the dependent variable, $In\mathrm{Y}_B$, and the independent variables, InX_{1B}, InX_{2B}, InX_{3B}, InX_{4B} and InX_{5B}. From the results outputs of the multiple linear correlation analysis on Data B (Now with Heteroscedasticity Remedied), presented in Table 3.30, it was evident that the dependent variable, $In\mathrm{Y}_B$, specifically has an average positive correlation with InX_{2B}; weak positive correlation with InX_{3B}; weak negative correlations with InX_{1B} and InX_{4B} (in that order of magnitude); and no correlation at all with InX_{5B}.

Finally, the predicted values of the dependent variable obtained using (4.6) as well as their reversed-transformed values are, respectively, presented in Columns 8 and 9 of Table 3.2 (see Appendix B). Also, Figure (4.4) presents the graph of these reversed-transformed (supposed real) predicted values superimposed with the graph of the original values of the dependent variables (see Appendix C).

5. Conclusion

The very essence of correcting for (otherwise referred to as " remedying") the presence of heteroscedasticity, where it exists, in regression modelling has been demonstrated in this study. In order to illustrate this expression, this study employed two different hypothetical data; namely, Data A (the Original) and Data B (the Original). The two datasets satisfied the normality, multicollinearity and autocorrelation assumptions, but could not satisfy the homoscedasticity assumption (that is, the existences of heteroscedasticity were established in the two datasets).

The Ordinary Least Square (OLS) method was used to estimate the multiple linear regression models for Data A (the Original) and Data B (the Original); which are presented in (4.1) and (4.3), respectively. The model established for Data A (the Original) is seen to be statistically significant (that is of good fit) with an R-square value of 0.976, an AIC value of 332.5929, and an SBC value of 347.2533. In a likely manner, the model established for Data B (the Original) is also statistically significant with an R-square value of 0.553, an AIC value of 69.89669, and an SBC value of 82.15499.

The Log-transformation was applied on the variables in the two different datasets (Data A (the Original) and Data B (the Original)) that showed the existences of heteroscedasticity. These transformations gave rise to new sets of data now referred to as, Data A (Now with Heteroscedasticity Remedied) and Data B (Now with Heteroscedasticity Remedied). These Log-transformed datasets equally satisfied the normality, multicollinearity and autocorrelation assumptions, and also satisfy the homoscedasticity assumption (that is, there are no existences of heteroscedasticity in the two datasets).

The estimated multiple linear regression models for Data A (Now with Heteroscedasticity Remedied) and Data B (Now with Heteroscedasticity Remedied) are as presented in (4.2) and (4.4), respectively. The model established for Data A (Now with Heteroscedasticity Remedied) is seen to be statistically significant (that is of good fit) with an R-square value of 0.986, an AIC value of -135.021, and an SBC value of -120.361. In a likely manner, the model established for Data B (Now with Heteroscedasticity Remedied) is also statistically significant with an R-square value of 0.624, an AIC value of -32.0801, and an SBC value of -19.8218.

The values of the R-square for Data A (Now with Heteroscedasticity Remedied) and Data B (Now with Heteroscedasticity Remedied) are, respectively, greater than the values of the R-square for Data A (the Original) and Data B (the Original). It could be seen that 0.986>0.976 and 0.624>0.553. Also, the values of the AIC and SBC for Data A (Now with Heteroscedasticity Remedied) and Data B (Now with Heteroscedasticity Remedied) are, respectively, lesser than the values of the AIC and SBC for Data A (the Original) and Data B (the Original). It could also be seen that -135.021<332.5929; -120.361<347.2533 and -32.0801<69.89669; -19.8218<82.15499.

Now, from the points of view of the values of the R-square, AIC and SBC, it is evident that the estimated regression models for Data A (Now with Heteroscedasticity Remedied) and Data B (Now with Heteroscedasticity Remedied) are better models when compared to the regression models for Data A (the Original) and Data B (the Original).

Appendix B

Table 3.2: The Hypothetical Data B

Y_B	X_{1B}	X_{2B}	X_{3B}	X_{4B}	X_{5B}	$\hat{Y}_B$	$In\hat{Y}_B$	Exp. $(In\hat{Y}_B)$
0.787	3	3	3	3	5	2.993	-0.107730	0.897870
0.293	8	30	8	8	5	4.793	0.772799	2.165821
1.710	3	6	6	6	5	2.849	0.243002	1.275071
0.203	4	4	4	12	5	-1.185	-1.337640	0.262465
0.806	8	7	6	5	5	2.116	-0.255490	0.774538
4.713	10	20	5	5	5	3.304	0.405826	1.500542
0.607	8	6	3	3	25	2.008	-0.129330	0.878680
9.107	6	24	4	4	25	5.143	1.266464	3.548284
9.210	4	10	12	4	25	5.753	1.487012	4.423855
1.365	16	12	8	4	25	1.659	0.232780	1.262103
4.554	3	10	8	8	25	2.983	0.763792	2.146400
0.293	8	3	3	3	25	1.543	-0.764950	0.465357
2.252	3	6	3	3	50	3.278	0.771959	2.164002
9.167	3	8	8	3	50	4.983	1.643878	5.175200
0.694	4	8	4	8	50	1.183	-0.039910	0.960874
0.379	5	2	2	2	50	2.313	-0.499960	0.606556
0.485	2	2	2	3	50	2.653	-0.144640	0.865330
3.345	10	15	3	3	50	2.755	0.596045	1.814927
0.208	15	6	2	3	50	-0.289	-0.837790	0.432664
0.201	15	6	2	3	75	-0.389	-0.794820	0.451665
0.329	10	4	3	3	75	0.95	-0.573030	0.563817
4.966	3	8	2	2	75	3.691	1.245389	3.474287
1.362	6	6	6	4	75	2.711	0.363074	1.437742
1.515	2	3	8	6	75	2.936	0.415015	1.514393
0.751	5	2	2	2	75	2.213	-0.456980	0.633193
1.568	4	8	4	8	100	0.983	0.033562	1.034131
1.203	2	4	4	12	100	-1.017	-0.435070	0.647218
0.806	9	7	6	5	100	1.462	-0.037350	0.963339
2.613	8	24	5	5	100	4.092	1.078896	2.941430
3.972	9	6	3	3	100	1.434	-0.081800	0.921460
7.107	4	28	4	2	100	6.975	2.611615	13.621030
6.213	2	10	6	2	100	5.291	2.503858	12.229590
0.694	2	10	4	8	50	2.041	0.749727	2.116422
1.379	9	13	2	2	50	2.922	0.720392	2.055239
2.485	6	8	2	3	50	2.487	0.199359	1.220620
3.345	13	9	3	3	50	1.003	-0.093820	0.910449
1.208	10	8	2	3	50	1.391	-0.231780	0.793123
0.401	16	9	2	3	75	-0.198	-0.477470	0.620348
2.329	9	6	3	3	75	1.534	-0.112290	0.893785
3.966	5	9	2	2	75	3.298	0.922260	2.514967
1.362	7	12	6	4	75	3.367	0.868586	2.383539
2.515	3	11	8	6	75	3.902	1.264245	3.540418
0.751	4	7	8	8	75	2.044	0.310370	1.363930
0.787	6	6	3	3	100	2.256	0.260417	1.297471
1.293	8	21	8	8	100	3.018	0.763276	2.145293
1.568	6	6	6	6	100	1.647	-0.024470	0.975830
1.203	2	4	4	12	100	-1.017	-0.435070	0.647218
0.806	9	7	6	5	100	1.462	-0.037350	0.963339
3.613	8	20	5	5	100	3.472	0.911707	2.488567
3.972	9	6	3	3	25	1.734	-0.228740	0.795533
8.107	4	26	4	2	25	6.965	2.396711	10.986980
7.213	2	10	6	2	25	5.591	2.356911	10.558290
1.365	14	12	8	4	25	2.207	0.345480	1.412668
3.345	10	15	3	3	50	2.755	0.596045	1.814927
0.208	15	6	2	3	50	-0.289	-0.837790	0.432664
0.201	15	6	2	3	75	-0.389	-0.794820	0.451665
0.329	10	4	3	3	75	0.950	-0.573030	0.563817
0.787	3	3	3	3	5	2.993	-0.107730	0.897870
0.293	8	30	8	8	5	4.793	0.772799	2.165821
1.710	3	6	6	6	5	2.849	0.243002	1.275071

Appendix C

Plots of the Y Values of the Hypothetical Data, their Predicted Y Values and the Predicted Y Values of the Hypothetical Data (Now with Heteroscedasticity Remedied)

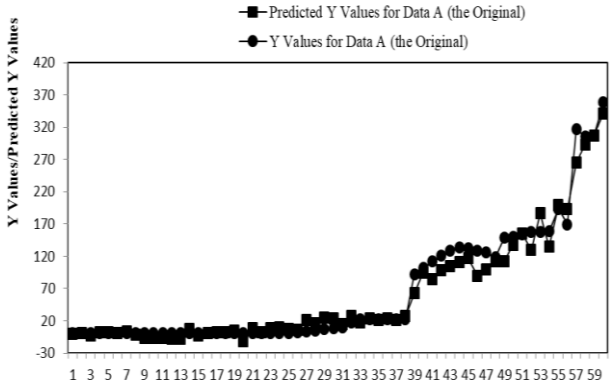

Fig.4.1: Plots of Y Values and the Predicted Y Values for Data A (the Original).

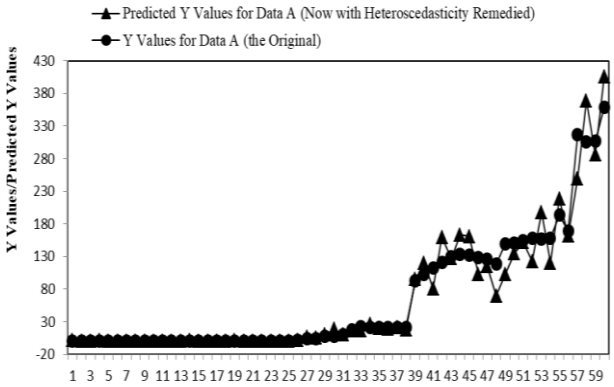

Fig. 4.2: Plots of Y Values for Data A (the Original) and the Predicted Y Values for Data A (Now with Heteroscedasticity Remedied).

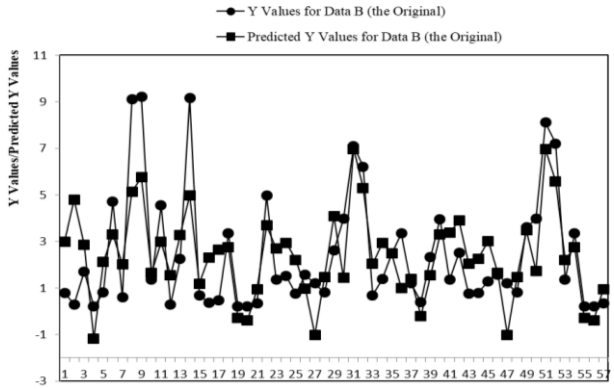

Fig. 4.3: Plots of Y Values and the Predicted Y Values for Data B (the Original).

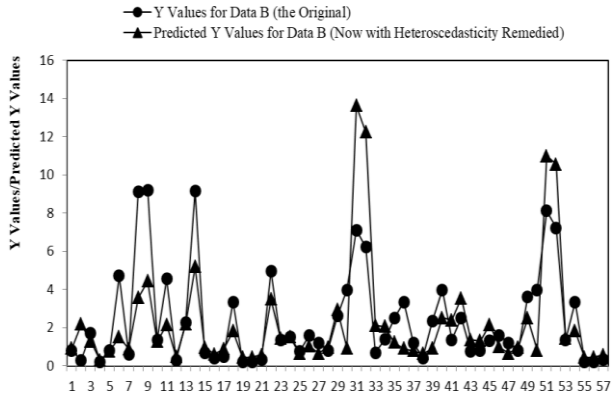

Fig.4.4: Plots of Y Values for Data A (the Original) and the Predicted Y Values for Data A (Now with Heteroscedasticity Remedied).

References

[1] A. Adepoju and P. O. Tayo. Regression Methods in the Presence of Heteroscedasticity and Outliers. *Academia Journal of Scientific Research, 5(12):* (2017); *776-783.*

[2] O. L. O. Astivia and B. D. Zumbo. Heteroskedasticity in Multiple Regression Analysis: What it is, How to Detect it and How to Solve it with Applications in R and SPSS. *Practical Assessment, Research and Evaluation, 24(1)* (2019). Available online: http://pare-online.net/getvn.asp?v=24&n=1

[3] Cohen, P., West, S. G. and Aiken, L. S. (2007). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Mahwah, NJ: Erlbaum.

[4] A. Field. *Discovering Statistics Using SPSS* (3rd ed.). Sage, London, UK, (2009).

[5] J. Fox. *Applied Regression Analysis, Linear Models and Related Methods*. Sage, London, UK, (1997).

[6] M.O. Gidigbi and M. Donga. Domestic, Foreign Direct Investment and Economic Growth Nexus in Selected African Countries. *AUDOE, 17(5):* (2021); *142-157.*

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

[8] J. F. Hair, R. E. Anderson, R. L. Tatham and W. C. Black. *Multivariate Data Analysis*, (3rd ed.). Macmillian, New York, U.S.A, (1995).

[9] A. F. Hayes and L. Cai. Using Heteroskedasticity-Consistent Standard Error Estimators in OLS Regression: An Introduction and Software Implementation. *Behaviour Research Methods, 39:* (2007); *709-722. https://doi.org/10.3758/BF03192961*.

[10] K. Jabłońska. Dealing with Heteroskedasticity within the Modelling of the Quality of Life of Older People. *Statistics in Transition New Series, 19(3):* (2018); *433-452. https://doi.org/10.21307/stattrans-2018-024.*

[11] A. Koutsoyiannis. *Theory of Econometrics* (7th ed.). Macmillian, London, United Kingdom, (1977). https://doi.org/10.1007/978-1-349-09546-9.

[12] M. H. Kutner, C. J. Nacthtsheim, J. Neter and Li Williams. *Applied Linear Statistical Models* (5th ed.). Mcgraw-Hill/Irwin, New York, U.S.A, (2005).

[13] M. T. Nwakuya and J. C. Nwabueze. Application of Box-Cox Transformation as a Corrective Measure to Heteroscedaticity Using an Economic Data. *African Journal of Mathematics and Statistics, 8(1):* (2018); *8-12.*

[14] S. C. Nwankwo. *Econometrics: A Practical Approach.* El' demak, Enugu, Nigeria, (2011).

[15] E. U. Ohaegbulem and V. C. Iheaka. The Impact of Macroeconomic Factors on Nigerian-Naira Exchange Rate Fluctuations (1981-2021). *Asian Journal of Probability and Statistics (AJPAS), 26(2):* (2024); *18-36. https://doi.org/10.9734/ajpas/2024/v26i2589*.

[16] J. O. Rawlings, S. G. Pantula and A. D. Dickey. *Applied Regression Analysis: A Research Tool* (2nd ed.). Springer-Verlag, New York, U.S.A, (1998). https://doi.org/10.1007/b98890.

[17] M. A. Stephens. The Anderson-Darling Statistic. (1979). https://www.google.com.ng/url?sa=source=web&rct=j&url=http://www.dtic.mil/dtic/tr/fulltext/u2/a079807.pdf&ved=2ahUKEwih5nKt9HeAhXK-MewKHeavA_MQFJAAegQLABAB&usg=AOvVaw3S-jPcRbLcJ9_Ovd7H8ONG

[18] F. Williams. Heteroskedasticity. https://www3.nd.edu/~rwilliam/ (2020)

[19] W. Yoo, M. Robert, B. Sejong, S. Karan, P. H. Qinghua and W. L. J. James. A Study of Effects of Multicollinearity in the Multivariable Analysis. *International Journal Applied Science Technology (IJAST), 4(5):* (2014) *9-19.*

[20] Q. M. Zhou, P. Z. K. Song and M. E. Thompson. Profiling Heteroscedasticity in Linear Regression Models. *The Canadian Journal of Statistics, 43(3):* (2017); *358-377. https://doi.org/10.1002/cjs.11252*.