Discriminating Between Ordinary Least Squares Estimation Method and Some Robust Estimation Regression Methods

BADMUS, NOFIU IDOWU AND OGUNDEJI, ROTIMI KAYODE
Department of Statistics
University of Lagos, Akoka,
NIGERIA

Abstract: - The lack of certain assumptions is common in ordinary least squares regression models whenever there is/are outliers and high leverage in the observations with an extreme value on a predictor variable. This could have a great effect on the estimate of regression coefficients. However, this research investigates the performance of the ordinary least squares estimator method and some robust regression methods which include: M-Huber, M-Bisquare, MM, and M-Hampel estimator methods. This study applies both methods to a secondary data set with 28 years (from 1900 to 2021) 200 meter races Summer Olympic Games with a response variable (sprint time) and three predictor variables (age, weight, and height) for illustration. Also, linearity, homoscedasticity, independence, and normality assumptions based on diagnostics regression like residual, normal Q-Q, scale-location, and cook's distance were checked. Then, the results obtained show that the robust regression methods are more efficient than the ordinary least square estimator method.

Key-Words: - Absolute Residual, Leverage, M-BiSquares, M-Hampel, M-Huber, Normal Q-Q, Outlier, Scale-Location

Received: July 2, 2022. Revised: August 29, 2023. Accepted: September 25, 2023. Published: October 31, 2023.

1 Introduction

In regression analysis, model fitting is always based on certain assumptions: linearity, homoscedasticity, independence, and normality. If the assumptions of the regression model, variables, and error terms are met, the application of the ordinary least squares approach in regression analysis works well. However, the OLS method of estimation becomes problematic when there are outliers (observation with large residual), high leverage points in which the explanatory variable turns away from its mean, and influence (the product of outlier and leverage) that can change the slope of the line or failure of the assumptions. This is because both good and bad leverage points, as well as vertical outliers, can have an impact on the model's residuals, coefficients, and standard errors [1]. Meanwhile, fitting a model requires regression diagnostics (an important tool) to evaluate the model assumptions and check whether or not there are observations with a large residual, outrageous, and undue influence on the analysis. Thereafter, we employ robustness checks to examine certain behaviors of regression coefficient estimates when modified by adding/removing regressors. Also, to reduce the impact of outliers the linearity assumption is still needed for proper inference using robust regression.

Numerous researchers have worked on this area in several ways: [2] investigated and defined vertical outliers as observations with outlying y-dimension values but not in predictor variables, impacting ordinary least squares estimation. [3] critically studied robust regression methods and fitted data that revealed the breakdown due to vertical outliers. However, he stated that robust methods effectively bound the influence of unusual observations, making them powerful statistical tools for identifying unusual observations. LTS performed more than ordinary least squares estimators when outliers were not removed. [4] discriminated between robust estimation methods to Ordinary least squares (OLSE) using a weak multi-co-linearity dataset. They found that OLSE is inefficient when outliers are introduced, with S-estimators performing better. Simulation studies showed OLSE is inactive with outliers.

[5] performed a simulation study to compare more than three estimation methods including the Ordinary Least Squares Method (OLSM), Least Absolute Deviations Method (LADM), M- Estimators (ME), TLS estimator, and Non-parametric Regression. They concluded that the Ordinary least squares method performed better without contamination. In the same vein, when outliers were introduced in the response and predictor variables, the method broke down. Then, non-parametric methods are highly performed with outliers in both X and Y dimensions. [6] made a comparative study between Huber ME and the OLSE, comparing robust regression methods such as the M, W, R estimator, least median of squares estimator, LTSE, and Re-weighted LSE. Among all, M-estimation is the most efficient method, minimizing standardized residuals and giving smaller weights to unusual observations. W-estimators depict the importance of each observation, and R-estimators compute data ranks. L-estimators compute linear combinations of order statistics, including LTS and LMS. In the end, the Huber ME outperforms the OLSE in both standard error (SE) and coefficient of determination (CoD). [7] developed the MM-estimator, the most efficient with a high breakdown point. It uses an S-estimator as an initial estimate, achieving high breakdown point properties. The robust estimator was weighted to ordinary least squares, showing no influence from outliers. [8] also compared Iteratively Reweighted Least Squares (OLSE) with other estimators, but found Huber has leverage points issues and OLSE performed poorly overall. [9] extensively discussed the mean in OLS and median in different ME methods. The estimator's performance was evaluated using a Monte Carlo simulation study and depends on the mean square error of the regression coefficient. Meanwhile, it was concluded based on the results that the proposed ordinary least square robust GA method performed better than the OLS MD method for sample sizes.

[10] stated that instead of the OLS method in the presence of outlier(s) or contamination or influential observation(s), it is better to use any of the robust regression methods such as ME (Huber, Hampel, and Bisquare), the LTSE, the S-estimation, and the MME method. However, he concluded that robust estimators had a positive effect on efficiency and reduced biasedness compared to the classical estimation method. [11] also explained in their work that the least squares method fails or underperforms in the case of outliers due to its unreliable results. Huber and Tukey bisquare, MM, and LTS estimator perform well even when there are outliers; and concluded that the M-Huber estimator is more efficient with outliers in the data set fitted. . [12] implemented some robust regression techniques that can help policymakers in formulating public policies. In [13], robust ridge and Liu estimators were made available in the literature. [14] proposed a new reweighted covariance based on a regression estimator by studying several robust estimators and [15] compared three main methods from different robust regression methods. The motivation for this study is it's useful for computer Scientists in understanding their data sets, analyzing, making decisions, and enhancing the efficiency of their algorithm based on statistical techniques considered.

2 Material and Methods

The regression model is defined as sections as here.

$$Y = X\beta + \varepsilon \tag{1}$$

where,

Y is the vector of the dependent variables in order $(n \times 1)$, X is the matrix of one or more predictor variables in order $\big(n \times (p+1)\big)$, β is the vector of regression coefficients in order $\big((p+1), 1\big)$, ε is the vector of the error term in order $(n \times 1)$ and p is the number of independent variables [9].

2.1 The Ordinary Least Squares Method (OLS)

The OLS has been in existence in literature for decades [21]. It is a common estimation method that has been commonly used for estimating linear models, and also, has the unbiased estimation property. Generally, the OLSE for regression coefficients can be obtained using the equations below:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2}$$

The normal density function is given by

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, \quad \mu > 0, \sigma > 0 \tag{3}$$

Also, the bivariate linear model is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (4)$$

In (3) since μ is a location parameter, then for the unbiased estimate $\mu = E(Y)$ and by taking $E(Y)$ in (3), we get

$$E(Y) = \beta_0 + \beta_1 X \quad (5)$$

By substituting (5) into (3), taking the logarithm, differentiating the outcome concerning β_0 and β_1, and equating to zero. Thereafter, making β_0 and β_1 the subject of the formula, the outcome can be estimated using the OLS method as follows:

$$\hat{\beta}_1 = \frac{n\sum yx_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum yx_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \quad (6)$$

and

$$\hat{\beta}_0 = \frac{\sum y \sum x_i^2 - \sum x_i y \sum x_i}{n\sum x_i^2 - (\sum x_i)^2} = \bar{y} - \hat{\beta}_1 \bar{x} \quad (7)$$

where,

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \text{ and } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

2.2 Robust Regression (RR)

RR is a choice method for least squares regression when data are infected with outliers or influential observations. In short, it is a form of weighted and reweighted least squares (WRLS) regression and can be used instead of LS regression. Furthermore, some high-leverage data points might surface in fitting an LS regression, and these data points might not be errors from data entry, either because they are from different populations. With this, there is no concrete reason to remove those points from the analysis. RR can be employed as an alternative method due to its ability to accommodate all the data points and treat all of them equally in OLS regression. In the same vein, RR weighs the observations differently based on how well-behaved these observations are.

This study examines the impact of outliers, leverage points, non-normality, and infections on classical LSE in linear regression analysis. Robust methods like ME with Huber and bi-square weighting due to their high standard in estimating LS regression. ME defines a weight function, where the weights depend on the residuals and vice versa.

Suppose that U is a diagonal matrix representing the weight function defined as:

$$U_{ii}(\omega_i) = \frac{\varphi(\omega_i)}{\omega_i} = \frac{\varphi\left[\frac{y_i - \sum_{i=1}^{v} x_{ij}\beta}{\hat{\sigma}}\right]}{\left[\frac{y_i - \sum_{i=1}^{v} x_{ij}\beta}{\hat{\sigma}}\right]}$$

The estimated equations for the model parameters:

$$\sum_{i=1}^{n} X_{ij}\, \varphi\left(\frac{y_i - \sum_{i=1}^{v} x_{ij}\beta}{\hat{\sigma}}\right) = 0\,;\, j = 1,2,\dots,p \quad (8)$$

where $\varphi(\omega_i) = \acute{p}(\omega_i)$ represents the influence function. Meanwhile, (8) is written in terms of the weighted function as follows:

$$\sum_{i=1}^{n} X_{ij}\, U_i \left[\frac{y_i - \sum_{i=1}^{v} x_{ij}\beta}{\hat{\sigma}}\right] = 0 \quad (9)$$

Also, the solution of the estimated (9) can be obtained by reweighted OLS iteratively Reweighted Least Squares (IRLS) as follows:

$$\hat{\beta}^t = (X' U^{t-1} X)^{-1} X' U^{t-1} Y \quad (10)$$

$\hat{\beta}^0$ are mostly represented by the OLS estimators as the initial estimates of the regression coefficients.

By applying the ME method using Huber, there are some steps to follow. These are:

a. By obtaining the initial estimations of the regression coefficients by one of the estimation methods as OLS method.

b. Determine residual value error term (e_i)

c. Compute the median (mn) of the error term (e_i)

d. Calculate the median MD = $|e_i - mn|$

e. Estimate the scale parameter σ by computing $\hat{\sigma}$ as follows:

$$\hat{\sigma} = \frac{MD}{0.6745}$$

f. Calculate ω_i, where, $\omega_i = e_i/\hat{\sigma}_i$

g. Calculate the diagonal values of weighted matrix W that are defined in (i)

h. Calculate $\hat{\beta}^{H2}$ using the weighted least squares (WLS) method as:

$$\hat{\beta}^{H2} = = (\mathrm{X}' U_{i-1}\mathrm{X})^{-1} X' U_{i-1} Y$$

i Repeat steps b – h to obtain a convergent value of $\hat{\beta}_i^{H2}$. [9]

Table 1. The Weight Function of some of the Estimators

Estimator	**Weight Function** $\boldsymbol{\varphi(u) = w(u)/u}$
Least Square	1
MM	$\begin{cases}\left[1-\left(\frac{u_i}{4.685}\right)^2\right]^2, & \lvert u_i\rvert \le 4.685 \\ 0, & \lvert u_i\rvert > 4.685\end{cases}$
Huber $e > 0$	$\begin{cases}1 & for\ \lvert u\rvert < e \\ \frac{e}{\lvert u\rvert} & for\ \lvert u\rvert \ge e\end{cases}$
Hampel $e, f, g > 0$	$\begin{cases}1 & for \quad \lvert u\rvert < e \\ \frac{e}{\lvert u\rvert} & for \quad e \le \lvert u\rvert < f \\ e\frac{g/\lvert u\rvert - 1}{g-f} & for\ f \le \lvert u\rvert \le g \\ 0 & otherwise\end{cases}$
Bisquare $e > 0$	$\begin{cases}\left[1-\left(\frac{u}{e}\right)^2\right]^2, & for\ \lvert u\rvert \le e \\ 0, & for\ \lvert u\rvert > e\end{cases}$

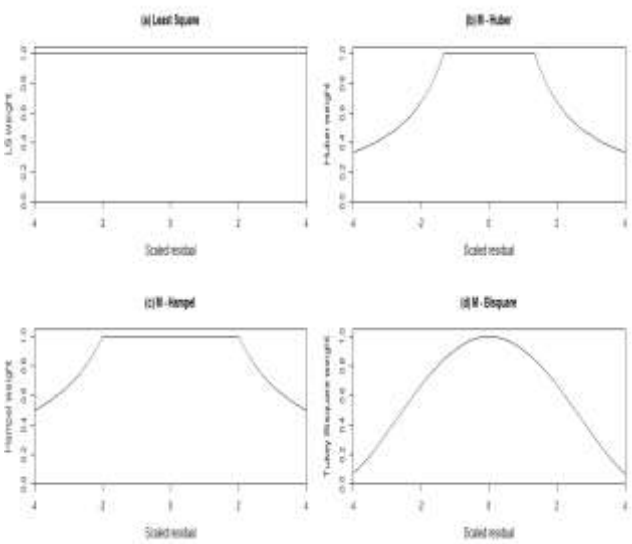

Fig. 1: Plots of the Weight Functions of the Table 1

3 Data Analysis

Here, the secondary data used for the illustration is extracted from a sports journal [16]. The variables are age, weight, height (predictor variable), and sprint time (response variable) of 28 Olympic game winners from 1900 - 2021. It has 28 data points; age and weight are used to predict their sprint time.

Table 2. Description of Variables

Variable	**Code**	**Description**
Sprint	Y = Sp	Time spent by each winner of the 200m Race Summer Olympic game
Age	X1 = Ag	Age of each 200m Race winner of the Summer Olympic game
Weight	X2 = Wg	Weight of each 200m Race winner of the Summer Olympic game
Height	X3 = Hg	Height of each 200m Race winner of the Summer Olympic game

Source: [12]

3.1 The Multiple Regression Model

In regression analysis, a regression model with more than one regressor variable is known as a multiple regression model. Researchers in the literature have discussed extensively on major assumptions of the

multiple regression model [17]. In this study, sprint time($= y_i$) (response variable), age $(= x_1)$, weight $(= x_2)$ and height $(= x_3)$ (predictor variable). The model is given by

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i \qquad (11)$$

where, β_0 is the intercept and ε_i is the error term, and it is believed that the distribution of y_i follows the ε_i.

3.2 The Diagnostics Plot of the Sprint Time.

Fitting regression model, the main tool for examining the fit is the residuals. In Figure 1 below there are 4 diagnostic plots of sprint time to depict if the data set conforms to the assumptions of linear regression or deviates/violates any.

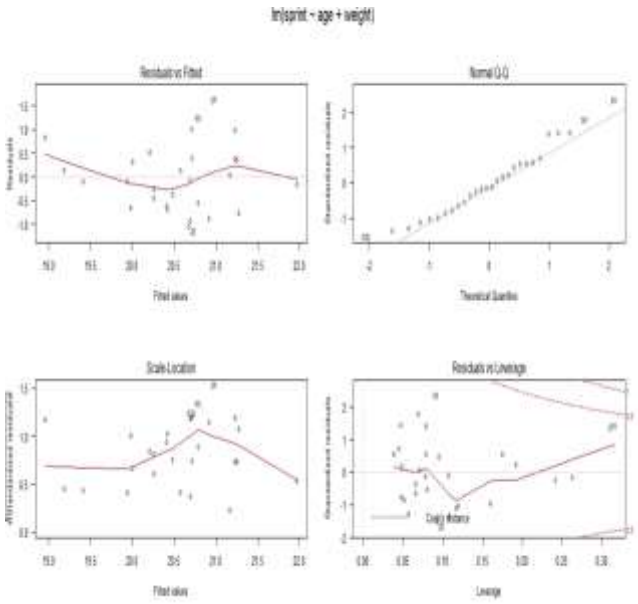

Fig. 2: The Diagnostic Plots of Residual, Normal Q-Q, Scale-Location and Leverage.

Table 3. The Coefficient, Standard Error (.) of the Estimators for Real Data Set with Normal Errors

Method	Intercept	$\widehat{\beta}_{Age}$	$\widehat{\beta}_{Weight}$	$\widehat{\beta}_{Height}$
OLS-Est	26.878(6.284)	-0.058(0.060)	-0.063(0.037)	-0.122(4.482)
Huber-Est	24.895(6.609)	-0.056(0.063)	-0.070(0.039)	1.216(4.714)
Bisquare-Est	25.526(6.668)	-0.059(0.064)	-0.068(0.039)	0.806(4.756)
Hampel-Est	26.424(6.379)	-0.059(0.061)	-0.065(0.037)	0.208(4.549)
MM-Est	25.600(6.641)	-0.059(0.063)	-0.068(0.039)	0.755(4.737)

Table 4. The Mean Square Error (MSE), Root Mean Square Error (RMSE) Coefficient of Determination (CD)

Method	MAP	RMSE	MAPE	MAD	CD
M-Huber-Est	1.0000	1.0000	1.0000	1.0000	0.7097
M-Bisquare Est	0.9898	0.9696	0.9999	1.0000	0.7176
M-Hampel Est	1.0000	1.0000	1.0000	1.0000	0.7082
MM-Est	0.3357	0.5794	-0.0011	1.27e-16	0.7352

4 Results and Discussion

4.1 Results

Figure 1 explains the various ways in which the $\varphi(u)$ weigh the scaled residuals. It is obvious that the least squares estimator only assigns weight one to all observations, but M-estimators' weight functions reduced weights at the tails. This implies that the OLS method cannot handle unusually large residuals as the robust M-estimators will have control over it. In a nutshell, M-estimators are more robust in governing heavy-tailed error distributions and non-constant error variance [18]. Meanwhile, the nature of outliers determines the kind of weight function to be selected and used by the researchers [19]. Then, Fig. 1 is critically studied, and the differences between M-estimators can be well understood.

Fig. 2, illustrates the diagnostic plot of residuals vs fitted values. Residuals are measured as follows:

$$residual = actual\ Value\ (y) - predicted\ Value\ (\bar{y})$$

The purpose of the plot of residuals vs predicted values is to check the level of assumption of linearity and homoscedasticity, the normal QQ plot is used to determine normality assumption in observations. Also, a scale-location plot is useful for checking the assumption of homoscedasticity. While Cook's distance is a measure of the influence of each observation on the regression coefficients. Generally, in these plots, it can be easily identified that observations in years: 3, 5, and 28 are possibly problematic to the model. This is one of the major reasons some robust M-estimator methods that can handle outliers, leverage, and influence observations than the OLS method, are employed.

Table 1 contains the weight function of OLS, MM, Huber, Hampel, and Tukey Bisquare estimators with their mathematical expressions, and the plots are shown in Fig 1. Table 2 consists of the description of variables in the multiple regression model and the codes used in R software to generate the output in Table 3. Therefore, Table 3 summarizes the results (the coefficients) of the multiple regression analysis performed on the real data set. Based on the outcome, none of the predictors has a statistically significant contribution from the OLS method, but only variable height generated a positive value in all robust estimators considered in the study and it is statistically significant. Although other variables (age and weight) are not, this means that height has a positive impact on y (sprint time). That is, the time each winner finished the race.

4.2 Model Selection Criteria

In this study, the following estimators were used and considered when outliers appear in the data set, MSE, RMSE, MAPE, MAD, and CD were used as model selection criteria. Comparisons of the model were made according to the identity that the lower the value for MSE, RMSE, MAPE, MAD, and CD the more valuable a model can fit the data [20].

However, Table 4, narrates how closely each model is to fit the data. It was determined that the selection criteria were investigated, and revealed that the MME has a smaller value in terms of MSE, RMSE, and MAD, the M-Bisquare estimator also has a smaller value in terms of MAPE and also M-Hampel estimator has a smaller value in terms of CD. It is therefore recommended to use the MM-estimator to estimate the sprint time of the winners of the 200m race of the summer Olympic game and any related observations.

4.3 Conclusion

This study demonstrated and determined the effective performances of MM, M-Huber, M-Hampel, and M-Bisquare estimators due to the failure of the OLSE in basic linear regression assumptions. Firstly, the model predictions in the data set when there are: outliers, high leverage, and influence were obtained. Secondly, applying a correct estimator to analyze variables of interest will yield appropriate, accurate, and reliable results (s). Thirdly, we concluded that the model performance of the MME is preferable with outliers, high leverage, and influence in the data set. Finally, for further study, we suggest that more robust regression methods should be considered using both real observations and simulation data respectively. This could allow Computer scientists to have an impact on the study by generating an algorithm for simulating data to illustrate the estimators.

.

Acknowledgement:
The authors thank those who helped in the proofreading of this work.

References:

[1] Ogundeji, R. K, Onyeka-Ubaka, J. N. and Yinusa, E. **(2022).** Comparative Study of Bayesian and Ordinary Least Squares Approaches. *Unilag Journal of Mathematics*

and Applications. ISSN: 2805 3966. Vol 2 (1) pp. 60 – 73.

[2] Verardi, V. and Croux, C. (2009). Robust regression in Stata. *The Stata Journal*, 3:439–453.

[3] Fox, J. and Weisberg, S. (2010). An appendix to an r companion to applied regression second edition. 1–17.

[4] Cetin, M. and Toka, O. (2011). The comparison of s-estimator and m-estimators in linear regression. *Gazi University Journal of Science*, 24(4):747–752.

[5] AL-Noor, H. N. and Mohammad, A. (2013). Model of robust regression with parametric and nonparametric methods. *Mathematical Theory and Modeling*, 3:27–39.

[6] Bhar, L. (2014). Robust regression. http://www.iasri.res.in/ebook/EBADAT/3- Diagnostics

[7] Yohai, V.J. (1987). High breakdown-point and high-efficiency robust estimates for regression. The Annals of Statistics 1987; 15: 642- 656.

[8] Ruppert, D., Street, J. O., and Carroll, R. J. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, 42:152–154.

[9] Ismail, I. M. and Rasheed, H. A. (2021). Robust Regression Methods/ a Comparison Study. Turkish Journal of Computer and Mathematics Education; Vol 12(14), 2939 – 2949.

[10] Morrison, T. S. (2021). Comparing Various Robust Estimation Techniques in Regression Analysis. Graduate Theses, Dissertations, and Other Capstone Projects. Minnesota State University, Mankato. 1 – 54.

[11] Tirink, C and Onder, H. (2022). Comparison of M, MM, and LTS estimators in linear regression in the presence of outlier. Turkish Journal of Veterinary & Animal Sciences. Vol. 46(3)., 420 – 428 https://doi.org/10.55730/1300-0128.4212

[12] Khan, D. M., Yaqoob, A., Zubair, S., Khan, M. A., Ahmad, Z and Alamri, O. A. (2021). Applications of Robust Regression Techniques: An Economic Approach. Hindawi, Mathematical Problems in Engineering. Vol. 2021, 1 – 9. https://doi.org/10.1155/2021/6525079

[13] Adegoke, A. S., Adewuyi, E., Ayinde, K and Lukman, A. F. (2016). A Comparative Study of Some Robust Ridge and Liu Estimators. Science World Journal, Vol. 11(4), 16 – 20. www.scienceworldjournal.org

[14] Lakshmi, R. and Sajesh, T. A. (2023). Empirical Study on Robust Regression Estimators and Their Performance. RT& A, Vol. 18(2), 466 – 478.

[15] Shafiq, M., Amir, W. M and Zafakali, N. S. (2017). Algorithm for Comparison of Robust Regression Methods in Multiple Linear Regression by Weighting Least Square Regression (SAS). Journal of Modern Applied Statistical Methods, Vol. 16(2), 490 – 505. Doi.10.22237/jmasm/1509496020.

[16] Ugofotha, M. O., Ogwumu, O. D and Nwaokolo, M. A. (2023). A Sport Model for Predicting the Sprint Time for the Winning of 200m Race of a Summer Olympic Games. Asian Journal of Pure and Applied Mathematics, Vol. 5(1), 73 – 87.

[17] Montgomery, D. C, Peck, E. A and Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. https://books.google.com/books?hl=en&lr=&id=tCIgEAAAQBAJ&oi=fnd&pg=PP13&dq=douglas+c+montgomery+an+introduction+to+regression&ots=lfseWyl1Sn&sig=WqfV-rA-XmrqPL-qCFzek_0gLPk.

[18] Andersen, R. (2008). Modern Methods for Robust Regression. Thousand Oaks: SAGE Publications.

[19] Rousseeuw P.J. and Leroy A.M., 1987, Robust Regression and Outlier Detection, John Wiley, New York, 202. [24]Fox J., "Robust Regression", An R and S-PLUS Companion to Applied Regression, http://cran.r-project.org/.

[20] Tatliyer, A. (2020). The effects of raising type on performances of some data mining algorithms in lambs. Journal of Agriculture and Nature; 23 (3): 772-780.

[21] Stigler, S. M. (1981). Gauss and the invention of Least Squares. The Annals of Statistics, Vol. 9(3), 465 – 474.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Badmus, Nofiu Idowu initiated the idea, worked on the introduction, material, and methods, and did the data analysis using the R software package to generate outcomes in Figures 1 and 2, and Tables 4 and 5. While
Ogundeji, Rotimi K. explained the outcomes in Figures 1 and 2 and Tables 4 and 5 in Section 4 and made the concluding remarks. But they both wrote the abstract and generally proofread the paper.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself
No funding was received for conducting this study.

Conflict of Interest
The authors have no conflicts of interest to declare that are relevant to the content of this article.