International Journal of Development Mathematics Vol 1 Issue 4 Page No. 177 - 190

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

Handling Multicollinearity and Outliers: A Comparative Study of Some One and Two–Parameter Estimators Using Real-Life Data

Oyeleke K. Tayo[a*], Timothy O. Olatayo[a] and Biodun T. Efuwape[a]

[a]*Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Nigeria.*

ARTICLE INFO

Article history:
Received 07 July 2024
Received in revised form 15 September 2024
Accepted 19 October 2024

Keywords:
Multicollinearity, Outliers, traditional least square, Mean square error, estimator

MSC 2020 Subject classification:
62P05

ABSTRACT

It is evident that when data suffers the problem of multicollinearity, the traditional least square is incapacitated and unreliable. Hence, needs to use bias estimator such as Ridge estimator, Liu estimator among others. Also, presence of outliers is another treat and to tackle this challenge is the use of robust regression estimators which include M, MM, LTS, LMS, LAD, LQS and S estimators. However, the presence of the two anomalies may be inevitable. Several estimators have been combined to handle the problems simultaneously. Therefore, this study compared and contrasted some robust one and two-parameter estimators using some real-life data sets. Mean Square Error (MSE) was used as criterion to select the best estimator. Some of the robust estimators were found to be inconsistent in addressing the twin problems. However, across all the data set employed in the study, the results revealed that robust Modified Ridge Type (MRT) in M, MM and LTS did well using minimum MSE.

1. Introduction

Regression model plays a significant role in data analysis, as it is being used in various fields to analyze data. It serves as predictive modeling technique that examines the relationship that exists between two variables (dependent and independent variables). (Jege *et al.*, 2022). However, extreme values and multicollinearity pose serious problems for regression analysis, resulting in skewed estimates and untrustworthy conclusions. When predictor variables exhibit strong correlation, multicollinearity arises, leading to exaggerated variances of the predicted coefficients. Outliers, or extreme values, can have an unbalanced impact on the regression model and produce inaccurate findings. Also, regression analysis has long been known to have issues with multicollinearity. Farrar and Glauber (1967) highlighted the problems it produces as early as 1967 which include inflated standard errors and untrustworthy statistical tests. Many robust estimators have been proposed in order to lessen the effects of multicollinearity such as principal component regression Jolliffe, (1982), partial least squares regression Wold *et al*. (1984), and ridge regression (Hoerl and Kennard, 1970) have been proposed to handle multicollinearity. Hoerl and Kennard (1970) created ridge regression, which reduces variance and shrinks coefficients by adding a regularization component to the ordinary least squares (OLS) objective function. Ridge regression may still be sensitive to extreme values even with its efficacy. By converting the predictor variables into uncorrelated principal components, Principal Component Regression (PCR) resolves multicollinearity (Jolliffe, 1982). Furthermore, Liu estimator by Liu (1993), the ridge regression estimator and the Principal Component Regression (PCR) estimate were combined by Baye and Parker (1984) to develop the r–k class estimator. Similarly, the *r–d* class estimator was developed by Kaciranlar and Sakallioglu (2001) using PCR and the Liu estimator.

Similarly, other researchers have combined two or more estimators in order to circumvent the problem of multicollinearity in regression models. These include the two-parameter estimator by Ozkale and Kaciranlar (2007), the *k-d* class estimator by Sakallioglu and Kaciranlar (2008), Yang and Chang (2010) created a novel two-parameter estimator by merging the PCR estimator with the Unbiased Ridge Regression (URR) estimator of Crouse *et al*. (1995). Similarly, other researchers have combined two or more estimators in order to circumvent the problem of multicollinearity in regression models. These include the modified r –k class ridge regression (MCRR) estimate was developed by Batah *et al*. (2009) by merging the Unbiased Ridge Regression (URR) estimator by Crouse *et al*. (1995) with the PCR estimator, a new two parameter estimator by Yang and Chang (2010). Meanwhile, Mansson *et al*. (2018) worked on a few estimators for the ridge parameter of the multinomial logit model, evaluating the efficacy of various

* Corresponding author. Tel.: +2349164353019
E-mail address: kamoruoyeleke@gmail.com (Oyeleke, K. Tayo)
https://doi.org/10.62054/ijdm/0104.14

estimators taken into consideration using MSE as a yardstick. The DK estimator by Dawoud and Kibria (2020), Ahmad and Aslam (2020) introduced a novel two-parameter estimator termed the Modified novel Two-type parameter Estimator (MNTPE) when exogenous variables are correlated in a multiple regression model. Among other things, Lukman *et al*. (2020) also merged the modified ridge-type estimator and the PCR estimator. In the case of Alabi *et al*. (2014), they examined the tolerable sample size needed for traditional estimator to be used when multicollinearity exist among the regressors. Idowu *et al.* (2023) proposed the two-parameter version of Kibria-Lukman estimator in a bid to further deal with the problem of multicollinearity in linear regression analysis. In the same spirit, Idowu *et al*. (2023) developed a two-parameter estimator for correlated Regressors in Gamma Regression Model which is an expansion of linear regression model whereby Gamma Modified Two-Parameter (GMTP) estimator was proposed as a means of dealing with the problem of multicollinearity. Their study revealed that the newly proposed estimator outperformed other estimators such as Maximum Likelihood Estimator (MLE), Gamma Ridge Estiamtor (GRE), Gamma Liu Estimator (GLE) and Gamma Liu Type Estimator (GLTE) both theoretically and using Monte Carlo experiment. However, PCR among some other aforementioned estimators does not necessarily handle outliers effectively and coincidentally or naturally the two problems may be present in the data. Regression analysis becomes even more complex when extreme values are present. Robust regression techniques, such the least median of squares (LMS) and the least trimmed squares (LTS), were established by Rousseeuw and Leroy (1987) to reduce the impact of outliers. Although these techniques work well against outliers, multicollinearity may not be sufficiently addressed. Majeed *et al*. (2021) have suggested a robust M-Kibria-Lukman estimator, which is a relatively new invention that addresses extreme values and multicollinearity. In order to increase efficiency and robustness, Kibria (2003) presented a class of ridge estimators that utilize robust regression concepts. This work was expanded upon by Lukman *et al*. (2012), who introduced hybrid estimators that combine robust approaches and ridge regression's advantages. Simulation research and empirical applications have demonstrated the potential of these hybrid estimators, offering improved performance over traditional methods. The significance of maintaining equilibrium between bias and variance when multicollinearity and outliers are present was highlighted by Kibria and Lukman (2012). They proved that typical ridge regression and robust regression techniques are not as effective at achieving this balance as their suggested estimators. Further research, such that done by Özkale and Kibria (2017), has confirmed the usefulness of Kibria-Lukman estimators in more situations. Modified Ridge M-estimator by Hassan (2017) and Lukman *et al*. (2019) proposed a robust estimator that can handle three joint problems in regression analysis which include; multicollinearity, outliers, and autocorrelation. Dawoud and Abonazel (2021) proposed Robust Dawoud-Kibria and so on. More recently is Robust M-New Two Parameter estimator by Adejumo *et al*. (2023) as an alternative estimator that can simultaneously suppress the problem of multicollinearity and outliers in linear regression analysis. Hence, this work focuses on a comparative investigation of a few one and two-parameter estimators. The following robust regression estimators which include (M, MM, S, LTS, LMS, LAD and LQS) estimators were combined with the MRT-estimator. The goal is to assess these estimators' performance in terms of minimum Mean Squares Error (MSE) using some real-life data.

2.1 Materials and Methods

2.1.1 The Traditional Least Square Estimator

The traditional linear regression model also known as Ordinary Least Square (OLS) is given in equation (1) where X is an $n \times p$ known matrix independent variable with full rank, y is an $n \times 1$ vector response variable, $\gamma_{p\times 1}$ is the $p \times 1$ vector of unknown regression coefficients and U is the $n \times 1$ vector random error terms with $E(U)=0$ and $E(U^T U)=\sigma^2 I_n$ such that I_n is an identity matrix of $n \times n$.

$$Y = X\gamma_{p\times 1} + U \qquad (1)$$

If $M=(I,X)$ and $\gamma=(\gamma_0,\gamma_1^T)^T$ therefore, the Ordinary Least Squares (OLS) of $\gamma_{p\times 1}$ in (1) can be written as:

$$\gamma_{OLS} = G^{-1}X^T y \qquad (2)$$

where $G = X^T X$

The Canonical form of the Traditional Least Square Estimator

The canonical form of the model in (1) can be written as:

$$Y = Z\alpha + U \tag{3}$$

where $Z = XT, \alpha = T^T\gamma$ and T is the orthogonal matrix with columns that constitute the eigenvalues of G. Hence, $Z^T Z = T^T G T = \Gamma = diag(\eta_1, \eta_2, ..., \eta_p)$ such that $\eta_1 \le \eta_2 \le ..., \le \eta_p$ are the ordered *G*.

Therefore, the traditional least squares estimator of α can be written as:

$$\hat{\alpha}_{OLS} = \Gamma^{-1} Z^T y \tag{4}$$

2.1.2 Ridge Regression Estimator

Heorl and Kennard (1970) proposed Ridge regression estimator in other to deal with the problem of multicollinearity in regression analysis. The Ridge parameter was added to the G matrix to reduce the collinearity effect. The OLS estimator of α tends to have a large variance which was defined as;

$$\hat{\gamma}_{ORR} = (G + kI)^{-1} X^T y. \tag{5}$$

where *I* is the $p \times p$ identity matrix and $k = \frac{P\sigma^2}{\sum_{i=1}^{p} \hat{\alpha}_i^2}$ is the ridge parameter such that *p* is the number of explanatory variable, $\sigma^2 = \frac{\sum_{i=1}^{p} U^2}{n-p}$ is the Mean Square Error (MSE) and $\hat{\alpha}_i$ is an unbiased estimator of $\hat{\gamma}$.

2.1.3 Liu Regression Estimator

To overcome the problem of multicollinearity in the data sets, Liu (1993) proposed the Liu Estimator by combining the Stein Estimator with Ridge estimator to form;

$$\hat{\gamma}_L = (G + I)^{-1}(G + dI)\hat{\gamma}_{OLS}. \tag{6}$$

where $d = 1 - \sigma^2 \left[\frac{\sum_{i=1}^{p} \frac{1}{\eta_i(\eta_i + 1)}}{\sum_{i=1}^{p} \frac{\alpha_i^2}{(\eta_i + 1)^2}} \right]$ such that λ_i is the ith Eigen value G and $\hat{\alpha}_i = T^T \hat{\alpha}$

2.1.4 Kibria-Lukman Estimator

Kibria and Lukman (2020) proposed another one parameter estimator that can combat the problem of multicollinearity in linear regression model aside already existing ones. They defined the estimator as:

$$\hat{\gamma}_{KL} = (G + kI_p)^T (G - kI_p)\hat{\beta}_{OLS} \tag{7}$$

where $k = \frac{\hat{\sigma}^2}{2\hat{\alpha}_i^2 + (\hat{\sigma}^2 / \eta_i)}$ and $\hat{\gamma}_{OLS}$ is the OLS estimator.

2.1.5 Modified Ridge Type (MRT) Estimator

Lukman *et al.* (2019) proposed the modified ridge-type estimator (MRT) as an alternative estimator to handle the problem of multicollinearity in linear regression model as it can be expressed in (8).

$$\hat{\gamma}^{MRT}(k,d) = A_{k,d}\hat{\beta} = (G + k(1+d))^{-1} X^T y. \tag{8}$$

where $A_{k,d} = (G + k(1+d)I)^{-1} G, k \succ 0$ and $0 \prec d \prec 1$.

They expressed *k* and *d* as follows:

$$k = k_{HMP}^{MRT} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}(1+d)\hat{\alpha}_i^2},$$

$$d = \hat{d}_{MRT} = \frac{p}{\sum_{i=1}^{p}\frac{1}{\hat{d}}} \text{ such that } \hat{d} = \min\left(\frac{\hat{\sigma}^2}{\alpha_i^2}\right)$$

They affirmed that, the estimator can give the same results with the OLS and the ridge estimator majorly when $k = 0$ and $d = 0$.

2.1.6 Dawoud-Kibria Estimator

As an alternative method of reducing the impact of multicollinearity in linear regression model, Dawoud and Kibria (2020) proposed a new biased estimator denoted as; $\hat{\gamma}_{DK}$ and defined as:

$$\hat{\gamma}_{DK} = (G + k(1+d)I_p)^{-1}(G - k(1+d)I_p)\hat{\gamma}_{OLS}. \tag{9}$$

such that; $k = \hat{k}_{\min MD} = \min\left(\frac{\hat{\sigma}^2}{(1+\hat{d})\left(\frac{\hat{\sigma}^2}{\eta_i} + (2\alpha_i^2)\right)}\right)_{i=1}^{p}$ and $d = \hat{d} = \min\left(\frac{\hat{\sigma}^2}{\alpha_i^2}\right)$.

2.2 Robust Regression Estimators

An iterative procedures that is designed to handle the problem of extreme observation(s) including influential points in the data sets and as well minimize their impact over the regression parameters is refers to as Robust regression (Zaman *et al*. 2001). In regression analysis, the main aim of robust estimation is to obtain dependable estimates or inferences for unknown parameters when there are anomalies such as outliers (Khan *et al*. 2021). Hence, the inferences obtained from robust estimation are more reliable because of their insensitivity to outliers either big or small. Some of the different kinds of robust regression are hereby discussed.

2.2.1 M-Estimator

Huber, (1964) proposed M-estimator as a robust method of dealing with outliers in regression analysis. Instead of minimizing the sum of squared errors, it minimizes the estimator $\hat{\gamma}$. It is an extension of the maximum likelihood estimate method. (Yuliana and Susanti, 2008; Susanti *et al.* (2014)). The estimator is defined as in (10):

$$\hat{\gamma}_M = \min \sum_{i=1}^{n} \rho\left(\frac{y_i - x_{ij}\hat{\gamma}}{s}\right). \tag{10}$$

where *s* is a robust estimate of scale expressed as $s = \frac{median|u_i - median(u_i)|}{0.6745}$.

2.2.2 MM-Estimator

` One of the most commonly robust regression techniques in the presence of extreme values is MM-estimator developed by Yohai (1987). In this technique, more than one M-estimation procedures were used to obtain the final estimates. The major aim of MM-estimator is to obtain parameters that have a high breakdown value and very efficient. The estimator is defined as:

$$\hat{\gamma}_{MM} = \sum \rho \left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\hat{\gamma}_j}{s_{MM}} \right) x_{ij} \quad . \tag{11}$$

where s_{MM} is the standard deviation obtained from the residual of *S*-estimation and ρ is a Tukey's biweight function. (Susanti *et al.*, 2014).

2.2.3 Least Median Squares (LMS) Estimator

As an alternative to the Ordinary Least Squares Estimator (OLSE), Rousseeuw (1984) proposed LMS since, median is not usually affected by outliers. Instead of the mean value in OLS, the median value is used. The Least Median Squares estimator can be expressed as:

$$\hat{\gamma}_{LMS} = \min{}_{\beta}[med(u_i^2)] \tag{12}$$

where, e_i is the residual from the OLS.

2.2.4 Least Trimmed Square (LTS) Estimator

As a highly efficient estimator to the Least Median Square (LMS), Rousseeuw (1984) proposed (LTS) estimator which is defined as;

$$\hat{\gamma}_{LTS} = \min\left(\sum_{i=1}^{h} (u_i^2)_{i=n} \right). \tag{13}$$

where $(u^2)_{1:n} \le (u^2)_{2:n} \le \ldots \le (u^2)_{n:n}$ are the ordered squared errors from the least to the highest and $h = n(1-\alpha)+1$ recommended by Rousseeuw and Leory (1987) such that α is the trimmed percentage.

2.2.5 *S*-Estimator

A high breakdown estimator which is capable of downsizing the dispersion of error terms in regression analysis is *S*-estimator developed by Rousseeuw and Yohai (1987). They defined the estimator by:

$$\hat{\gamma}_S = \min \hat{\sigma}_S (u_1, \ldots, u_n) . \tag{14}$$

where, $\hat{\sigma}_s = \sqrt{\frac{1}{nk}\sum_{i=1}^{n} w_i u_i^2}$ with $K = 0.199$ and initial estimate $\hat{\sigma}_s = \frac{median|u_i - median(u_i)|}{0.6745}$

2.2.5 Least Absolute Deviation (LAD)

The Least Absolute Deviation (LAD) also known as Least Absolute Error (LAE) was firstly introduced by Boscovich in 1757 (Birkes and Dodge, 1993). In the LAD, the coefficients are chosen so that the sum of the absolute deviation of the error terms is minimized. The mathematical expression of the estimator is defined as in equation (15):

$$\hat{\gamma}_{LAD} = \min \sum_{i=1}^{n} |y_i - x_i\hat{\gamma}| . \tag{15}$$

This method is very useful when the distribution of the residual does not follow a normal distribution. Also, LAD is robust when the contamination is in the y-direction. (Ahmed and Maha, 2016).

2.2.7 Least Quartile of Squares (LQS) Estimator

Least Quartile of Square (LQS) estimator was introduced by Rousseeuw (1984) and it is the generalization of Least Median Squares. It is based on the idea of choosing different quartiles. Also, LQS was built based on the proportion of observations equal to *q* such that a *1-q* proportion of observations is considered an extreme observation.

The estimator is defined as:

$$\hat{\gamma}_{LQS} \in \arg\min_{\beta} \left| r_{(q)} \right| \quad (16)$$

where $r_{(q)}$ denotes the residual corresponding to the qth ordered absolute residual:

$$\left| r_{(1)} \right| \leq \left| r_{(2)} \right| \leq \ldots \leq \left| r_{(n)} \right|.$$

2.3 Some Robust One and Two Ridge Regression Estimators

2.3.1 Ridge-M Regression

The ridge regression estimator was introduced by Hoerl and Kennard (1993) since OLSE is inefficient in the presence of multicollinearity. This was done by introducing a biasing parameter k into the design matrix of G. Also, it was noted that the Ridge Regression estimator is always affected by outliers in the y-direction, which led Silvapulle (1991) to propose robust ridge regression, defined as:

$$\hat{\gamma}_k^M = \left(kG^{-1} + I_p\right)^{-1} \hat{\gamma}_M \quad (17)$$

where $\hat{\gamma}_M = \min_{\gamma} \sum_{i=1}^{n} \theta\left(\frac{u_i}{k}\right)$, such that $\hat{\gamma}_M$ is the M-estimator, $k \geq 0$ and $u_i = y_i - x_i^T \hat{\gamma}_M$.

The generalized robust Ridge Estimator is defined as:

$$\hat{\gamma}_R^K = \left(kG^{-1} + I_p\right)^{-1} \hat{\gamma}_R, \quad (18)$$

where $\hat{\gamma}_R$ is each of the robust regression estimator (M, MM, S, LTS, LMS, LAD and LQS)

2.3.1 Robust Liu Estimator

The Liu estimator is another biased estimator to handle the problem of multicollinearity in a linear regression model. It was introduced by Liu (1993), which can be expressed as:

$$\hat{\gamma}_L = (G + I)^{-1}(X^T y + d\hat{\gamma}). \quad (19)$$

where $0<d<1$ and d are the biasing parameters. Meanwhile, the Liu estimator has been noted to be affected by the extreme values, especially in the y-direction; this led Arslan and Billor (2000) to propose its robust version, which can be defined as follows:

$$\hat{\gamma}_M^d = \left(G + I_p\right)^{-1} \left(G + dI_p\right) \hat{\gamma}_M. \quad (20)$$

The generalized robust Liu estimator is given as:

$$\hat{\gamma}_R^d = \left(G + I_p\right)^{-1} \left(G + dI_p\right) \hat{\gamma}_R. \quad (21)$$

2.3.2 Robust Kibria-Lukman Estimator

As an alternative to the one-biasing parameter estimator aside Ridge and Liu estimator, Kibria and Lukman (2020) proposed K-L estimator, defined as:

$$\hat{\gamma}_{KL} = \left(G + kI_p\right)^{-1} (G - kI_p) \hat{\gamma}_{OLS}. \quad (22)$$

The robust version of the K-L estimator when there are outliers in the y-direction was just recently proposed by [21], defined as:

$$\hat{\gamma}_M^{KL} = \left(G + kI_p\right)^{-1} \left(G - kI_p\right) \hat{\gamma}. \quad (23)$$

The generalized robust KL-estimator is defined as:

$$\gamma_R^{KL} = \left(G + kI_p\right)^{-1} (G - kI_p) \hat{\gamma}_R. \quad (24)$$

2.4 Robust Two-Parameter Estimator

In a bid to curtail the effect of multicollinearity in linear regression analysis, Ozkale and Kaciranlar (2007) came up with Two-parameter estimator. They defined the estimator as;

$$\hat{\gamma}_{TP} = \left(G + kI_p\right)^{-1} \left(G + kdI_p\right) \hat{\gamma}_{OLS}. \quad (25)$$

However, due to the sensitivity of the two-parameter estimator to outliers in the y-direction Awwad *et al* (2022) proposed a robust version expressed as:

$$\hat{\gamma}_{TP}^{M} = G\left(G + kI_p\right)^{-1}\left(G + kdI_p\right)\hat{\gamma} \quad (26)$$

The generalized robust TP estimator can be expressed as:

$$\hat{\gamma}_{R}^{TP} = G\left(G + kI_p\right)^{-1}\left(G + kdI_p\right)\hat{\gamma}_R \quad (27)$$

2.5 Robust Dawoud-Kibria Estimator

As an alternative, Dawoud and Abonazel (2021) proposed the DK estimator, which was noted to outperform others under some conditions and simulation studies. The estimator can be expressed as:

$$\hat{\gamma}_{DK} = \hat{\gamma}(DK) = \left(G + k(1+d)I_p\right)^{-1}\left(G - k(1+d)I_p\right)\hat{\gamma}_{OLS}. \quad (28)$$

to be already existing estimators that can deal with the problem of multicollinearity,
since the presence of extreme observations in the response variable direction has been noted to influence the performance of the Dawoud-Kibria estimator, Hence, to combat this problem, Dawoud and Abonazel (2021) proceeded and proposed a robust version of DK estimator by introducing $\hat{\beta}_M$ instead of $\hat{\gamma}_{OLS}$ used in the DK estimator. They defined the estimator as:

$$\hat{\gamma}_M(DK) = \left(G + k(1+d)I_p\right)^{-1}\left(G - k(1+d)I_p\right)\hat{\gamma}_M. \quad (29)$$

The generalized robust DK estimator is given as:

$$\hat{\gamma}_R(DK) = \left(G + k(1+d)I_p\right)^{-1}\left(G - k(1+d)I_p\right)\hat{\gamma}_R. \quad (30)$$

2.6 Generalized Robust Modified Ridge Type (RMRT) Estimator.

Lukman *et al.* (2020) proposed robust M-Modified Ridge Type estimator but when there are outliers in the y-direction. The generalized version of robust-MRT is defined as:

$$\hat{\gamma}_{(k,d)}^{RMRT} = \left(G + k(1+d)I_p\right)^{-1}G\gamma_R. \quad (31)$$

where γ_R is the individual robust regression estimators (MM, LTS, LMS, LAD, S and LQS), whereas *k* and *d* are the estimated biasing parameters for the robust MRT.

Assume that $\hat{\gamma}_R \sim N(0,1)$. This implies that $\hat{\gamma}_R$ follows a normal distribution with mean zero, variance equal one and covariance matrix equals A^2G^{-1}. When $\sqrt{n}\left(\hat{\gamma}_R^2 - \gamma\right) \sim N(0, A^2\Gamma^{-1})$, hence the assumption sustains most importantly for practical use, where $A^2 = \frac{c_0^2 E\left[\omega^2(\varepsilon / c_0)\right]}{E\left[\omega'(\varepsilon / c_0)\right]^2}$. Such that c_0 is the scale estimate.

Also, the unbiased estimator $\gamma_{iR} = \hat{\gamma}_{iR}$ that is $E(\gamma_{iR}) = \hat{\gamma}_{iR}$ and the unbiased estimator $\Omega_{ii} = A^2 / G_i$ such that

$$A^2 = \frac{c^2(n-p)^{-1}\sum_{i=1}^{n}\left[\omega(\varepsilon_i / c)\right]^2}{\sum_{i=1}^{n}\left[\frac{1}{n}\omega'(\varepsilon_i / c)\right]^2}.$$

The Harmonic mean of the biasing parameters for RMRT is as follows:

$$\hat{k} = \hat{k}_{RHM}^{MRT} = \frac{p\hat{A}^2}{\sum_{i=1}^{p}(1+d)\gamma_{iR}^2}. \quad (32)$$

where $d = \hat{d}_{RHM}^{MRT} = \frac{p}{\sum_{i=1}^{p}\frac{1}{\hat{d}}}$ such that $\hat{d} = 0 < \hat{d} < 1$

2.7 Investigation of the Estimators

In order to examine the performance of all the estimators conserved in this study, the MSE of all the estimators were used as yardstick. Their MSEs were estimated as follows:

$$MSE(\hat{\alpha}_R(k)) = \sum_{i=1}^{p}\frac{\eta_i^2\Omega_{ii}}{(\eta_i+k)^2} + \sum_{i=1}^{p}\frac{k^2\hat{\alpha}_i^2}{(\eta_i+k)^2} \quad (33)$$

Liu M-estimator has the MSE of:

$$MSE(\hat{\alpha}_R(d)) = \sum_{i=1}^{p}\frac{(\eta_i+d)^2\Omega_{ii}}{(\eta_i+1)^2} + \sum_{i=1}^{p}\frac{(1-d)\hat{\alpha}_i^2}{(\eta_i+1)^2} \quad (34)$$

Robust Kibria-Lukman has the following MSE:

$$MSE(\hat{\alpha}_R(KL)) = \sum_{i=1}^{p}\frac{(\eta_i+k)^2\Omega_{ii} + 4k^2\hat{\gamma}_i^2\eta_i}{\eta_i(\eta_i+k)^2} \quad (35)$$

MSE of robust Dawoud-Kibria is expressed as:

$$MSE(\hat{\alpha}_R(DK)) = \sum_{i=1}^{p}\frac{(\eta_i-k(1+d))^2\Omega_{ii}}{(\eta_i+k(1+d))^2} + \sum_{i=1}^{p}\frac{4k^2(1+d)^2\hat{\gamma}_i^2}{(\eta_i+k(1+d))^2} \quad (36)$$

MSE of MRT is given as:

$$MSE(\hat{\alpha}(MRT)) = \frac{\eta_i\sigma^2 + k^2(1+d)^2\hat{\alpha}_i^2}{(\eta_i+k(1+d))^2} \quad (37)$$

MSE of Robust MRT is given as in (32):

$$MSE(\hat{\alpha}_R^{MRT}(MRT)) = \frac{\eta_i\Omega_{ii} + k^2(1+d)^2\hat{\alpha}_i^2}{(\eta_i+k(1+d))^2} \quad (38)$$

The value of k and d used in this study are one used by Yasin and Murat (2016) which 0.0012 and 0.5 respectively.

3. Results

3.1 Data Description

Performances of the estimators were examined using two real-life data sets, the description of the data sets are as follows:

3.1.1 First data set: Longly data

The first data set is an economic data used by several authors such as Jahufer and Jianbao (2009), Jahufer (2013), Yasin and Murat (2016), Ullah *et al*. (2013), Kashif *et al.* (2019) and Lukman and Ayinde (2018), equation (39) is the regression model for the data.

$$y = \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4 + \gamma_5 x_5 + \gamma_6 x_6 \quad (39)$$

where *y* is the total derived employment, x_1 is the gross national product implicit price deflator, x_2 is the gross national product, x_3 is unemployment, x_4 is the size of armed forces, x_5 is the non-institutional population 14 years of age and over and x_6 is the time. Meanwhile, according to Walker and Birch, (1988) affirmed that the scaled Condition Number (CN) of the data is 43.275. Likewise, Liu (1993) claimed that the data suffered from the problem of multicollinearity in the values of the Variance Inflation Factor (VIF), which were estimated to be 128.29, 103.43, and 70.87. Likewise, Midi affirmed the claim of Liu (1993) and spotted the presence of outliers in the data.

3.1.2 Second data set: Hussein and Abdalla

The second data set used is the data adopted by Hussein and Abdalla (2012). The linear model below is the regression model for the data set.

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{40}$$

where y is the product value in the manufacturing sector, x_1 is the value of the imported intermediate, x_2 represents the imported capital commodities and x_3 indicates the value of imported raw materials. Hussein and Abdalla (2012) claimed that the data suffers the problem of multicollinearity as the values of Variance Inflation Factor (VIF) which was estimated to be (128.26, 103.43, and 70.87). Likewise, Lukman, *et al*. (2014) affirmed the claim of Hussein and Abdalla (2012) and spotted the presence of outliers in the data.

The value of parameter *k* and *d* used in this study are the one used by Yasin and Murat (2016), Ullah *et al.* (2013) and Kibria and Lukman (2020) which was computed as 0.0012 and 0.5 respectively.

3.1.3 Third data set: Pasha and Shah data

The regression model for the third data set is:

$$y = \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4 + \gamma_5 x_5 \tag{41}$$

The data was adopted from the study of Pasha and Shah (2004) and used by Kibria and Shipra (2016). y is the number of persons employed in (million), x_1 is land cultivated in million hectares, x_2 is the inflation rate in percentage, x_3 is the number of establishments, x_4 is the population in million meanwhile x_5 is the literacy rate in percentage. Diagnosis of the data revealed that the data suffers the problem of multicollinearity except x_2 with Variance Inflation Factor (VIF) 32.14, 2.50, 26.27, 71.03 and 22.70 respectively. Also, cases 1, 22 and 27 were identified as influential points indicating that the data set also suffers the problem of outliers.

Table 1: MSE of all the Estimators using Longley data

	Robust-M			**Robust-MM**			**Robust-LTS**	
Est	**MSE**	**Rank**	**Est**	**MSE**	**Rank**	**Est**	**MSE**	**Rank**
MRT_M	**3.711595**	**1**	**MRT_MM**	**3.297306**	**1**	**MRT_LTS**	**5.84645**	**1**
RE_M	3.715881	2	RE_MM	3.301336	2	RE_LTS	5.852522	2
KL_M	1033.002	3	KL_MM	967.7889	3	KL_LTS	1474.428	3
DK_M	747209.3	4	DK_MM	695469.5	4	DK_LTS	1079083	4
M	848073.6	5	MM	795134.5	5	LTS	1208647	5
LE_M	4.42E+14	6	LE_M	4.14E+14	6	LE_LTS	6.29E+14	6

	Robust-LMS			**Robust-LAD**			**Robust-LQS**	
Est	**MSE**	**Rank**	**Est**	**MSE**	**Rank**	**Est**	**MSE**	**Rank**
MRT_LMS	**29.79715**	**1**	**MRT_LAD**	**5.277665**	**1**	**MRT_LQS**	**6.846347**	**1**
RE_LMS	32.15782	2	RE_LAD	5.282754	2	RE_LQS	6.852588	2
KL_LMS	531407.3	3	KL_LAD	1242.789	3	KL_LQS	1531.184	3
DK_LMS	4.42E+08	4	DK_LAD	1017337	4	DK_LQS	1252269	4
LMS	4.43E+08	5	LAD	1017598	5	LQS	1252591	5
LE_LMS	2.30E+17	6	LE_LAD	5.30E+14	6	LE_LQS	6.52E+14	6

	Robust-	

	S	
Est	**MSE**	**Rank**
MRT_S	**6.074775**	**1**
RE_S	6.080729	2
KL_S	1452.106	3
DK_S	1188988	4
S	1189293	5
LE_S	6.19E+14	6

Source: Author's computations

Table 2: MSE of all the Estimators using Hussein and Abdalla data

Robust-M			**Robust-MM**			**Robust-LTS**		
Est	**MSE**	**Rank**	**Est**	**MSE**	**Rank**	**Est**	**MSE**	**Rank**
DK_M	**1910.324**	**1**	**DK_MM**	**3111.204**	**1**	**DK_LTS**	**1065.013**	**1**
MRT_M	2211.919	2	MRT_MM	4849.813	2	MRT_LTS	3615.173	2
RE_M	2212.154	3	RE_MM	4850.327	3	RE_LTS	3615.556	3
M	2212.623	4	MM	4851.356	4	LTS	3616.323	4
KL_M	99846.65	5	KL_MM	218921.9	5	KL_LTS	163189.9	5
LE_M	1.02E+20	6	LE_MM	2.24E+20	6	LE_LTS	1.67E+20	6

Robust-LMS			**Robust-LAD**			**Robust-LQS**		
Est	**MSE**	**Rank**	**Est**	**MSE**	**Rank**	**Est**	**MSE**	**Rank**
DK_LMS	**3594.858**	**1**	**DK_LAD**	**3040.12**	**1**	**DK_LQS**	**6327.969**	**1**
MRT_LMS	3596.003	2	MRT_LAD	3041.088	2	MRT_LQS	6329.983	2
RE_LMS	3596.384	3	RE_LAD	3041.41	3	RE_LQS	6330.654	3
LMS	3597.147	4	LAD	3042.056	4	LQS	6331.997	4
KL_LMS	162324.6	5	KL_LAD	137275.6	5	KL_LQS	285737.2	5
LE_LMS	1.66E+20	6	LE_LAD	1.40E+20	6	LE_LQS	2.92E+20	6

Robust-S		
Est	**MSE**	**Rank**
DK_S	**3632.8**	**1**
MRT_S	3633.957	2
RE_S	3634.342	3
S	3635.113	4
KL_S	164037.8	5
LE_S	1.68E+20	6

Source: Author's computations

Table 3: MSE of all the Estimators using Pasha and Shah data

	Robust-M			**Robust-MM**			**Robust-LMS**			**Robust-LQS**	
Est	**MSE**	**Rank**	**Est**	**MSE**	**Rank**	**Est**	**MSE**	**Rank**	**Est**	**MSE**	**Rank**
MRT_M	**68.9264**	**1**	**MRT_MM**	**78.43**	**1**	**MRT_LMS**	**78.0765**	**1**	**MRT_LQS**	**85.8117**	**1**
RE_M	100.8	2	RE_M	114.7	2	RE_LM	114.2	2	RE_LQ	125.5	2

	24		M	3		S	09		S	23	
M	304.163	3	MM	346.1	3	LMS	344.541	3	LQS	378.675	3
RL_M	84724.6	4	RL_MM	96406	4	RL_LMS	15615.5	4	RL_LQS	17667.9	4
LE_M	1.50E+17	5	LE_MM	2.00E+17	5	LE_LMS	1.70E+17	5	LE_LQS	1.80E+17	5

	Robust-LTS			**Robust-LAD**			**Robust-S**	
Est	**MSE**	**Rank**	**Est**	**MSE**	**Rank**	**Est**	**MSE**	**Rank**
MRT_LTS	**78.0436**	**1**	**MRT_LAD**	**76.658**	**1**	**MRT_S**	**77.8087**	**1**
RE_LTS	114.16	2	RE_LAD	112.13	2	RE_S	113.817	2
LTS	344.396	3	LAD	338.28	3	S	343.358	3
RL_LTS	95931.4	4	RL_LAD	18254	4	RL_S	19518.3	4
LE_LTS	1.70E+17	5	LE_LAD	2.00E+17	5	LE_S	1.60E+17	5

Source: Author's computations

3.2 Discussion

From the data description, as it has been revealed in section 3.1.1, 3.1.2 and 3.1.3 that the data sets really suffer the problem of both multicollinearity and outliers as their VIFs is greater than 10 except x_2 in section 3.1.3. Hence, from Table 1, it was observed that MRT_MM outperformed others has it has minimum MSE and the closest rival is RE_MM. In the same vein RE_M, MRT_LTS and RE_LTS in this order also did well. Meanwhile, the results from Table 2 revealed that DK_LTS is better when compared with other estimators whereas, MRT_M competes favourable with little margin. Applying the estimators to real-life data of Pasha and Shah (2004) as the results is shown in Table 3, with the use of robust estimators, the problems have been mitigated. However, it was observed that MRT_M is the most efficient robust estimator among other robust estimators considered in the study as it has the minimum MSE.

4. Conclusion

When data is contaminated with multicollinearity and extreme values, it is evident that OLS is underperformed. In fact some of the robust estimators were found to be inconsistent in addressing the twin problem in term of their MSE. However, MRT_M, MRT_MM and MRT_LTS performed better. Also, RE_MM and DK_LTS did well. It was observed that M, MM, LTS, LMS, LAD, S, KL_M, KL_MM, LE_M, LE_MM among others could not perform well when both multicollinearity and extreme values occur in the data set.

Conflict of interest: The Authors hereby declare that there is no conflict of interest.

References

Adejumo, T. J., Ayinde, K., Akomolafe, A. A., Makinde, O. S. and Ajiboye, A. S. (2023). Robust-M new two-parameter estimator for linear regression models: Simulations and applications. *African Scientific Reports. DOI:10.46481/asr.2023.2.3.138*

Ahmed, M. G. and Maha, E. Q. (2016). Regression Estimation in the presence of outliers: A comparative study. *International Journal of probability and statistics,5(3).65 -72. DOI:10.5923 /j.ijps.20160503.01.*

Ahmad, S. and Aslam, M. (2020). Another proposal about the new two-parameter estimator for linear regression model with correlated regressors. *Communication in Statistics-simulation and computation.* https://doi.org/10.1080/03610918.2019.1705975.

Alabi, O.O., Olatayo, T.O. and Afobi, F. R. (2014). Empirical Determination of the Tolerable sample size for OLS Estimator in the presence of Multicollinearity. *Applied Mathematics 5, 1870- 1877. Doi: 10.4236/am.2014.513180.*

Arslan, O., and Billor. N. Robust Liu estimator for regression based on an M-estimator. J*ournal Applied Statistics. 27 (1): 39 -47* (2000).

Awwad, F. A., Dawoud, I. and Abonazel, M. R. (2022). "Development of robust Ozkale Kaciranlar and Yang- Chang estimators for regression models in the presence of multicollinearity and outliers". *Concurr Comput Prac Exp. e6779 34 (2022) https://doi:10.1002/cpe.6779.*

Batah, F. M., Ozkale, M. R. and Core, S. D. (2009). Combining Unbiased Ridge and Principal component regression estimators. *Communication statistics. Theory Methods. 38, 2201 – 2209.*

Birkes, D. and Dodge, Y. D. (1993). *Alternative methods of regression*, Wiley, New York.

Crouse, R., Jin, C. and Hanumara, R. (1995). Unbiased Ridge estimation with prior information and ridge trace. *Communication statistics. Theory Methods. 24. 2341 – 2354.*

Dawoud, I. and Abonazel, M. R. (2021). Robust Dawoud-Kibria estimator for handling multicollinearity and outliers in the linear regression model. *Journal of Statistical Computation and Simulation. DOI: 10.1080/00949655.2021.1945063.*

Dawoud, I. and Kibria, B. M. G. (2020). A new Biased Estimator to combat the multicollinearity of the Gaussian linear Regression model. *Stats. 3, 526 – 541. Doi: 10.3390/stats 3040033.*

Farrar, D. E., and Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. The Review of Economics and Statistics, 92-107.

Hassan, E. A., (2017). Modified Ridge M-Estimator for Linear Regression Model with Multicollinearity and Outliers. *Communication in Statistics and Computation. DOI: 10.1080/03610918.2017.1310231*.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Huber, P. J. (1964). Robust Estimation of a location parameter. *The Annals of Mathematical Statistics, 35, 73 - 101.*

Hussein, Y. A. and Abdalla, A. A. (2012). Generalized Two stages Ridge Regression Estimator for Multicollinearity and Autocorrelated errors. *Canadian Journal on Science and Engineering Mathematics, 3(3), 79 - 85.*

Idowu, J. I., Fashoranbaku, A. S. and Ayinde, K. (2023). A two-parameter estimator for correlated Regressors in gamma regression model. *Science World Journal Vol. 18(No 4). Doi:* https://dx.doi.org/10.4314/swj.v18i4.8

Idowu, J. I., Owolabi, A. T., Oladapo, O. J. Ayinde, K. Oshuporu, O. A. and Alao, A. N. (2023). Mitigating

Multicollinearity in Linear Regression Model with Two-Parameter Kibria-Lukman Estimators. *WSEAS Transactions*

Jegede, S. L., Lukman, A. F., Ayinde, K. and Odeniyi, K. A. (2022). Jacknife Kibria Lukman M-Estimator: Simulation and Application, *Journal of the Nigerian Society of Physical Sciences 4, pp 251-264. Doi: 10.46481/jnsps.2022.664*

Jolliffe, I. T. (1982). A note on the use of principal components in regression. Journal of the Royal Statistical Society: Series C (Applied Statistics), 31(3), 300-303.

Kaciranlar, S. and Sakallioglu. (2001). Combining the Liu estimator and the principal component regression estimator. *Communication statistics. Theory Methods. 30. 2699- 2705.*

Khan, D. M., Yaqoob, A., Zubair, S., Khan, M. A., Ahmad, Z. and Alamri, O. A. (2021). Applications of Robust Regression Techniques: An Econometric Approach. *Mathematical Problem in Engineering Vol. 202. Doi.org/10.1155/2021/6525079.*

Kibria, B. G. (2003). Performance of some new ridge regression estimators. Communications in Statistics – Simulation and Computation, 32(2), 419-435.

Kibria, B. M.G. and Shipra, B. (2016). Some Ridge Regression Estimators and Their Performances. *Journal of Modern Applied Statistical Methods, 15(1), 206 – 231.*

Kibria, B. M. and Lukman, A. F. (2020). A new Ridge-Type Estimator for the linear Regression model. *Simulations and Applications, Hindawi scientifica Vol. 2020. https://doi.org/10.1155/20209758378.*

Liu, K. (1993). A new class of biased estimate in linear regression. *Journal of Communications in statistics. Theory and Methods. 22:2, 393 – 402. Doi:10.1080/03610929308831027.*

Lukman, A. F., Adewuyi, E. Oladejo, N. and Olukayode, A. (2019). Modified Almost Unbiased Two-parameter Estimator in linear regression model. I.OP. Conf. series; *Material science and Engineering 640 (2019).012119. doi: 10.1088/1757. -899X640/012119.*

Lukman, A. F., Arowolo, O. and Ayinde, K. (2014). Some Robust Ridge Regression for for Handling Multicollinearity and Outliers. *International Journal of Sciences. Basic and Applied Research (IJSBAR), 16(2), 192 -202.*

Lukman, A. F., Ayinde, K., Aladeitan, B. and Bamidele, R. (2020). An unbiased estimator with prior information. *Arab Journal of Basic and Applied Sciences. 27:1, 45 – 55. Doi:10.1080/25765299.2019.1706799.*

Lukman, A. F., Kibria, B. G., and Saleh, A. M. (2012). Robust ridge regression estimators: Some comparisons. *Journal of Applied Statistics,* 39(5), 987-1001.

Majid, A., Ahmad, S., Aslam, M. and Kashif, M. A. (2021) Robust Kibria-Lukman estimator for linear regression model to combat multicollinearity and outliers. *Concurrency and Computation: Practice and Experience. Doi.org/10.1002/cpe.7533*.

Manson, K., Shukur, G. and Kibria, B. M. G. (2018). Performance of some ridge regression estimators for the multinomial logit model. *Communications in statistics - theory and methods, 47:12, 2795 – 2804. Doi: 10.1080/03610926.2013.784996.*

Ozkale, M. R. and Kaciranlar, S. (2007). The restricted and unrestricted two-parameter estimators. *Communication Statistics. Theory. Meth. 36, 2707 – 2725.*

Özkale, M. R., and Kibria, B. G. (2017). Some new ridge regression estimators and their performances. Communications in Statistics - *Theory and Methods*, 46(3), 1501-1518.

Pasha, G. R. and Shah.M. A. A. (2004). Application of Ridge regression to Multicollinearity data. *Journal of Research Science, (Sci), 97 – 106.*

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association, 79, 871-880.*

Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley and Sons.

Sakallioglu, S. and Kaciranlar, S. (2008). A new biased estimator based on ridge estimation. *Stat. Papers. 49, 669-689.*

Silvapulle, M. J. (1991). Robust ridge regression based on an M-estimator. *Australian Journal of Statistics: 33; 319 – 333.*

Susanti, Y., Hasil, P., Sri, S. H. andTwenty, L. (2014). M-Estimation, S Estimation and MM Estimation in Robust Regression. *International Journal of pure and Applied Mathematics 91, (3), 349 – 360.*

Ullah, M. A., Pasha, G. R. and Aslam, M. (2013). Assessing Influence on the Liu Estimators in Linear Regression Models. *Communications in Statistics – Theory and Methods, 42(17), 3100 – 3116.*

Walker, E. and Birch, J. B. (1988). Influence Measures in Ridge Regression. *Technometrics, 30(2), 221 – 227.*

Wold, S., Martens, H., and Wold, H. (1984). The multivariate calibration problem in chemistry solved by the PLS method. Proceedings of the Conference on Matrix Pencils, 286-293.

Yang, H. and Chang, X. (2010). A new two-parameter estimator in linear regression model. Communication in *Statistics. Theory and Methods 39 (6) .923 – 934. Doi:10.1080/03610920902807911.*

Yasin, A. and Murat, E. (2016). Influence Diagnostics in Two Parameter Ridge Regression. *Journal of Data Science, 14, 33 – 52.*

Yohai, V.J. (1987). High Breakdown-point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics. 15 (20): 642-656.*

Zaman, A., Rousseuw, P. J. and Orhan, M. (2001). "Econometric applications of high-breakdown robust regression techniques", *Economic Letters, 71, no.1,1-8.*