Diagnostic Influential Measures with New Two-Parameter Estimator in the presence of multicollinearity and extreme values: Simulations and Applications

TAIWO JOEL ADEJUMO[1*], KAYODE AYINDE[2], EMMANUEL TAIWO ADEWUYI[1]
CHRISTIANA TOYIN ADEJUMO[3]

[1]Department of Statistics, Ladoke Akintola University of Technology, Ogbomoso, NIGERIA
[2]Department of Mathematics and Statistics, Northwest Missouri State University, Maryville, Missouri, USA
[3]Department of Mathematics and Statistics, Bells University of Technology, Ota, Ogun State, NIGERIA.

Abstract: Identifying influential points in linear regression is vital for ensuring the validity of inferential conclusions. Traditional diagnostic measures, such as DFFITs (DFT), Cook's D (CKD), COVRATIO (CVR), Hadi's measure (HAD), Pena's statistic (PEN), and Atkinson statistic (ATK), are typically based on the Ordinary Least Squares (OLS) estimator, which assumes no violation of the basic linear regression assumptions. This study develops new diagnostic measures for these statistics using the New Two-Parameter (NTP) estimator to address multicollinearity. The study evaluated the performance of these measures through simulation studies with 1,000 replications under varying levels of multicollinearity, error variances, outlier percentages and magnitudes, and sample sizes. Results revealed that the newly proposed CVR measure with the NTP estimator achieved 100% detection of influential points and recorded the highest detection counts, outperforming all other measures. While traditional measures like CKD, PEN, and ATK based on OLS were effective only for small sample sizes in the absence of multicollinearity, their performance declined when multicollinearity was present. Conversely, CVRNTP consistently demonstrated superior performance when multicollinearity was mitigated. These findings suggest that the proposed CVRNTP is a robust tool for identifying influential points in datasets affected by multicollinearity. Real-life data applications further validated their performances.

Keywords: influential point, outliers, multicollinearity, Ordinary Least Squares, simulation studies

Received: March 22, 2024. Revised: October 16, 2024. Accepted: November 5, 2024. Published: December 31, 2024.

1 Introduction

Regression analysis is a critical tool in data analysis often used to model the relationship between a dependent variable and one or more independent variables. However, real-world data usually violate the assumptions of classical linear regression models. Two major challenges that often arise in regression analysis are multicollinearity and the presence of extreme values (outliers). Multicollinearity occurs when predictor variables are highly correlated, leading to instability in coefficient estimates. Outliers or extreme values, on the other hand, can distort the estimation process and reduce the reliability of inference on the model. The common Ordinary Least Squares (OLS) estimation methods are sensitive to both multicollinearity and extreme values, necessitating the development of robust estimators ([1]; [11]). In the same vein, in many economic problems, Ordinary Least Squares (OLS) have been the most common method of estimating regression parameters, but in compliance with some fundamental assumptions, such as; regressors must be measured without error, explanatory variables must be linearly independent, the error terms must be independently and identically normally distributed with mean zero and variance σ^2 among others. But, not in all cases do investigations reveal these assumptions to be satisfied. The cause could be the problem of high correlation among the predictors (multicollinearity) and the presence of extreme observations (outliers). To address the problem of multicollinearity in linear regression analysis, several biasing estimators have been developed in the literature, which include the ridge regression estimator by [24], [35]. [15] combined the principal component regression estimator with the two-parameter estimator proposed by [44], New Two-Parameter (NTP) by [62], the Modified two-parameter estimator proposed by [19] modified ridge-type estimator [40], *the* Kibria-Lukman estimator by [34], Dawoud-Kibria (DK) estimator by [18]. When exogenous variables are correlated in a multiple regression model, [2] proposed a new two-parameter estimator called Modified New Two-type parameter Estimator (MNTPE). [39] also combined the Principal Component Regression (PCR) estimator with the modified ridge–type estimator, but just to mention a few. Meanwhile, to deal with the problem of outliers, some of the already existing estimators are M – the estimator by [25], the S-estimator by [55], and MM-estimator by [65]. Others include the Least Median Squares (LMS) estimator proposed by [49] and the Least Trimmed Squares (LTS) estimator as a high efficiency to the LMS was also proposed by [49]. Also, the Least Winsorized Squares (LWS) estimate was proposed by [51]. Least Trimmed Mean (LTM) was proposed by [50]. By using these methods, it is possible to eliminate some of the data points, which in some cases need not be done, especially if that data point(s) is (are) important. [56]

To examine the effect of both good and bad leverage points on parameter estimation in regression analysis, [5] carried out a robust diagnostic analysis in this respect. The joint problems of outliers and multicollinearity in the data set may be inevitable, and if present, the least square estimator becomes incapacitated. Hence, there is a need for a robust estimator to deal with these anomalies. [43] through simulation studies, corroborated some estimators, and developed a robust estimator in addressing these two issues in regression analysis by combining the MM-estimator with the Ridge Regression estimator to form a robust Ridge regression based on the MM-estimator (RMM). [39] proposed a two-parameter ridge-type modified M-estimator when the linear regression model suffers both problems of multicollinearity and outliers. Other estimators to address these two problems include the Robust Ridge and Liu estimators [3], the Robust-M Dawoud-Kibria estimator by [17], and the Robust-M New Two Parameter by [1]. In the literature, measures for detecting influential points based on OLS have been proposed, and these include Cook's distance by [16], DFFITs by [16], and Welsch's distance [60]. [42] worked on the detection of single influential points in the Ordinary Least Squares (OLS) regression model, where the basic survey of the influential statistics of a single case, including exploratory analysis of all variables, were provided. Likewise, [28] proposed a deleting formula known as Modified Ridge Regression (MRR) in order to detect influential points in regression analysis. [27] proposed an appropriate deletion formula for the detection of influential points for the Liu estimator. [58] compared the performance of some influential measures, which include Cook's D, DFFITs, COVRATIO, Hadi's measure, and DFBETAS, in detecting influential points in the presence of multicollinearity at the choice of different Ridge parameters. [8] derived the generalized versions of DFFITs and Cook's D in two parameter ridge-type estimator and derived the approximate deletion formulas of the influential measures. [22] studied the influence of a few points using Pena's Statistic for the Ridge Regression. Meanwhile, Modified Pena's Statistics for the biased estimators were carried out by [38]. Another influential measure based on Pena's Statistic for one parameter, such as the Liu regression estimator, was proposed by [31]. [38] developed the generalized versions of Cook's D and DFFITs in two parameter Liu-Ridge estimator and derived the approximate deletion formula for the two influential measures. Also, [10] compared the performance of some robust estimators. Among the influential statistics and robust estimators considered in their study are Cook's D, Welsch-Kuh distance, DFFITs, DFBETAS, MM-estimator, Least Trimmed Square (LTS), and S-estimators respectively including OLS. Three real-life data sets were used to examine the performances of the estimators using Root Mean Square Error (RMSE) as a criterion, and they concluded that multiple high-leverage observations could be the source of multicollinearity in regression analysis. Hence, to tackle these challenges, [12] proposed another robust procedure for the parameter's estimation and revealed that the Diagnostic Robust Generalized Potentials (DRGP) for MM – estimator is the most efficient among the estimators considered which include Diagnostic Robust Generalized Potentials (DRGP-L) for L-estimator, Diagnostic Robust Generalized Potentials (DRGP-LTS) for LTS-estimator, Diagnostic Robust Generalized Potentials (DRGP-M) for M –estimator, Diagnostic Robust Generalized Potentials (DRGP-MM) for MM-estimator. [13] examined the effect of collinearity – influential points on data that has the problem of multicollinearity using the Monte Carlo experiment approach. In identifying prospective outlying cases in the Multiple Circular Regression Model (MCRM), [6] developed an outlier procedure using DFFITs statistic for circular cases.[32], through Monte Carlo experiment and application to real-life data, investigated the performances of some influential diagnostics for the Cox proportional hazards regression models and observed that the proposed Cook's distance for both standardized and adjusted residuals outperformed others in terms of influential points detection. To justify their claim, they applied their method to real-life data sets on bone marrow transplant Leukaemia.

This study aims to develop influential diagnostic tools for a new two-parameter estimator proposed by [62]. Effort is made to address the combined effects of multicollinearity and influential observations in ensuring more reliable estimates for the model parameters. Likewise, the approximate deletion formulas for Cook's D and DFFTITs for the estimator are derived. The study's significance lies in addressing a critical gap in regression analysis, where the simultaneous occurrence of multicollinearity and extreme values can distort model results. The development of these measures for the new two-parameter estimator enhances researchers to detect problematic observations that may adversely affect inferences on the linear regression model often used in economics, engineering, finance, and medical research.

2. Background of the study

The matrix form of linear regression model is written as:

$$y = X\theta_{p\times 1} + \ell$$

(1)

Where y is an $n\times 1$ vector dependent variable, X is an $n\times p$ standardized known matrix regressor, $\theta_{p\times 1}$ is the vector $p\times 1$ regression coefficients and ℓ is the $n\times 1$ vector of independent random error terms with $E(\ell)=0$ and $E(\ell^T\ell)=\sigma^2 I_n$. Such that I_n is an identity matrix of $n\times n$.

Suppose $W=(I,X)$ and $\theta=(\theta_0,\theta_1^T)^T$ therefore, the Ordinary Least Squares (OLS) of $\theta_{p\times 1}$ in (1) can be written as:

$$\hat{\theta}_{OLS} = (W^T W)^{-1} W^T y \qquad (2)$$

The residual can be expressed as:

$$\ell_i = y - \hat{y},$$
$$= y - W\hat{\theta}_{OLS},$$
$$= y - W(W^T W)^{-1} W^T y,$$
$$= \left(1 - W(W^T W)^{-1} W^T\right) y,$$
$$= \left(1 - H_{ii}\right) y$$

(3)

where $H_{ii} = W(W^T W)^{-1} W^T$ is the hat matrix also known as projection matrix.

2.1 Some Conventional Diagnostic Measures in Least Squares Estimator

In the literature, based on OLS estimator several conventional influential diagnostics have been developed, some of them are hereby discussed.

2.1.1 Cook's Distance Diagnostic Measure

Cook's distance is a measure of the distance between the least squares estimate based on all n-observations in $\hat{\theta}$ and the estimate derived when the *i^th^* observation is deleted. It is usually denoted by $\hat{\theta}_{(i)}$ and defined as:

$$D_i = \frac{\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^T \left(\hat{\theta}_{(i)} - \hat{\theta}\right)}{p\hat{\sigma}^2}$$

(4)

Where $\hat{\theta}_{(i)}$ and $\hat{\theta}$ respectively are the estimates when the *i*th observation is removed and when for full data. The measure is related to the distribution $f\ (p,\ n-p)$. Further algebraic expression of (4) can be given as:

$$D_i = \frac{t_i^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right).$$

(5)

where $h_{ii} = w(W^T W)^{-1} w^T$ is the *ith* diagonal elements of the hat matrix and where t_i is *i*th internally studentized residual which is defined as:

$$t_i = \frac{\ell_i}{\sqrt{\hat{\sigma}^2\left(1-h_{ii}\right)}}.$$

(6)

such that, $\ell_i = y_i - \hat{y}$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\ell_i^2}{n-p}$ which is the residual mean square.

[16] suggested that observations for which $D_i > 1$ warrants attention.

2.1.2 DFFITs Influential Measure

[61] proposed DFFITs diagnostic influential measure which is defined as the deletion influence of *ith* observation on the predicted or fitted value. Also DFFITs can be defined as the change in the predicted value for a point obtained when the data is full and *ith* data is deleted divided by the estimated standard deviation of the fit at that point. The statistic can be expressed as:

$$DFFITs_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}_{(i)}^2 h_{ii}}}, \quad i = 1, 2, 3, \ldots, n$$

(7)

Where $\hat{y}_i$ is the fitted value when data is full or complete, $\hat{y}_{(i)}$ is the fitted value when the *i*th observation is deleted, $\hat{\sigma}_{(i)}^2$ is the estimated mean square error *MSE* of $\hat{y}_{(i)}$. According to [38], equation (7) can also be expressed as:

$$DFFITs_i = \frac{w_i[\hat{\theta} - \hat{\theta}_{(i)}]}{S(w_i\hat{\theta})}.$$

(8)

where $S(w_i\hat{\theta})$ is an estimator of standard error of the fitted values, w_i is the ith row of the W matrix, $\hat{\theta}_{(i)}$ is the least squares estimator of θ when the ith case is deleted in fitting the linear regression model.

Further algebraic expression, equation (8) can be expressed as:

$$DFFITs_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \frac{\ell_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} ,$$

$$= \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i .$$

(9)

where $\hat{\sigma}^2$ is the estimate of σ, h_{ii} is the diagonal elements of the hat matrix H_{ii} and $t_i = \frac{\ell_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$ is the studentized residual.

[27] suggested that observations for which $/DFFITs_i$ $/>2\sqrt{\frac{p}{n}}$ warrants attention for large data sets and [20] claimed that if the absolute value of $DFFITs_i$ exceeds 1 for small to medium data sets, such observation is influential.

.

2.1.3 DFBETAS

DFBETAS is an influential measure that indicates how much the regression coefficient changes if *i*th observation were deleted. Such change is measured in terms of standard deviation units. The influential measure can be defined as:

$$DFBETAS_{ji} = \frac{\hat{\theta}_j - \hat{\theta}_{j(i)}}{\sqrt{\hat{\sigma}^2_{(i)} h_{ii}}}$$

(10)

where $\hat{\theta}_{j(i)}$ is the regression coefficient obtained when *i*th observation is removed, h_{ii} is the *i*th diagonal elements of the hat matrix H_{ii} and $\hat{\sigma}^2_{(i)}$ is the estimated mean square error *MSE* of $\hat{y}_{(i)}$ when the ith point is deleted.

2.1.4 Atkinson Diagnostic

[7] proposed modified Cook's distance denoted by (A_i) by replacing the variance $(\hat{\sigma}^2)$ of the full data used in Cook's distance with the variance $\hat{\sigma}^2_{(-i)}$ when the ith observation is deleted. The statistic is defined as:

$$A_i = \frac{\left(\hat{\theta} - \hat{\theta}_{(-i)}\right)^T W \left(\hat{\theta} - \hat{\theta}_{(-i)}\right)(n-p-1)}{(p+1)\hat{\sigma}^2_{(-i)}},$$

$$A_i = DFFITs_i \sqrt{\left(\frac{n-p-1)}{n}\right.} .$$

(11)

where $\hat{\theta}$ is the OLS estimator when data is complete, $\hat{\theta}_{(-i)}$ is the OLS estimator when the ith observation is deleted. The statistic can algebraically be expressed as:

$$A_i = \frac{t^2_{(-i)}}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right).$$

(12)

such that $t_{(-i)} = \frac{\ell_{(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)}(1-h_{ii})}}$, $\ell_{(-i)} = y_i - \hat{y}_{(-i)}$ and

$$\hat{\sigma}^2_{(-i)} = \frac{\sum_{i=1}^{n} \ell^2_{(-i)}}{n-p} .$$

The observation that its *PCDi* >1 identifies to be influential.

2.1.5 COVRATIO diagnostic influential measure

COVRATIO influential statistic is the ratio of the determinant of the covariance matrix when the *i*th observation is deleted to the determinant of the covariance matrix for the complete data [13]. The statistic is defined as:

$$COVRATIO_i = \left[\frac{\left|\left(W^T_{(i)} W_{(i)} S^2_{(i)}\right)^{-1}\right|}{\left|W^T_{(i)} W_{(i)} \hat{\sigma}^2\right|}\right] .$$

(13)

Equation (13) can be expressed as:

$$COVRATIO_i = \frac{\left[\frac{n-p-t_i^2}{n-p-1}\right]}{1-h_{ii}} .$$

(14)

where $t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$, *p* is number of independent variables and h_{ii} is the ith diagonal element of the hat matrix.

Alternatively, (13) can further be written as:

$$COVRATIO_i = \frac{(S^2_{(-i)})^{(p+1)}}{MS\hat{E}^{(p+1)}}\left(\frac{1}{1-h_{ii}}\right).$$

(15)

where $MS\hat{E}^{(p+1)} = \frac{SSE}{n-p+1}$, such that $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $S^2_{(-i)}$ is the variance when the ith observation is omitted.
An observation is considered influential if the following condition holds; $|COVRATIO_i| \geq \frac{3p}{n} + 1$.

2.1.6 Hadi's influential diagnostic measure

According to [23], Hadi's measure is an influential diagnostic that is based on the identification of influential points in both y and x directions. In this statistic, a single case deleted measure of leverage was introduced. Hence, the influence of the ith point can be measured by using notations in equation (12).

$$H^2_i = \frac{h_{ii}}{1-h_{ii}} + \left(\frac{p}{1-h_{ii}}\right)\left(\frac{d^2_i}{1-d^2_i}\right).$$
(16)

where, $d_i = \frac{\ell_i}{\sqrt{SSE}}$, p is the number of explanatory variable and h_{ii} is the ith diagonal elements of the hat matrix $W(W^TW)^{-1}W^T$.
The cut-off point for the statistic is given as $H^2_i = mean(H^2_i) + c\sqrt{\text{var}(H^2_i)}$, where c is an appropriately chosen constant to be 2 or 3 [21].

2.1.7 Pena's statistic

[46] proposed a new influential diagnostic measure apart from the existing ones proposed based on OLS such as Cook's D, DFFITs among others. The diagnostic measures the influence of a single observation by the rest of the data. He suggested that instead of examining the overall effect on the fitted value due to the elimination of one observation, how the deletion of each observation affects the prediction of a specific observation independently can be measured. The statistic is defined as follows:

$$S_i = \frac{1}{p\hat{\sigma}^2 h_{ii}} \sum_{j=1}^{n} \frac{h^2_{ji}\ell^2_j}{(1-h_{jj})^2}.$$
(17)

where h_{ii} is the ith diagonal element and h_{ji} is the jith element of the hat matrix $W(W^TW)^{-1}W^T$. In his study, he further revealed that (17) tends to follow a Gaussian distribution if the sample size and explanatory variables are large. He estimated the cut-off point for the statistic as $|S_i - median(S_i)| \geq 4.5MAD(S_i)$ and affirmed that a point is influential if the value of S_i is larger than $(S_i - E(S_i)/SD(S_i)$.

3. Influential Measures in New Two-Parameter Estimator

As a means of further dealing with the problem of multicollinearity in linear regression model, [62] proposed New Two-Parameter (NTP) estimator aside already existing ones. The estimator is defined as:

$$\hat{\theta}_{NTP} = \left(W^TW + I\right)^{-1}\left(W^TW + dI\right)\left(W^TW + kI\right)^{-1}W^TY.$$
(18)

where k and d are the biasing estimated parameters for the NTP estimator defined as:

$$k = \frac{\hat{\sigma}^2(\lambda_i + d) - (1-d)\lambda_i\hat{\alpha}^2_i}{(\lambda_i + 1)\hat{\alpha}^2_i},$$

(19)

$$d = \frac{\sum_{i=1}^{p}\left(\hat{\alpha}^2_i - \hat{\sigma}^2\right)\Big/(\lambda_i+1)^2}{\sum_{i=1}^{p}\left(\hat{\sigma}^2 + \lambda_i\hat{\alpha}^2_i\right)\Big/(\lambda_i+1)^2\lambda_i}.$$

(20)

$$0 < d \leq 1.$$

where λ_i is the ith Eigen value of W^TW, $\hat{\alpha}_i = Q^T\hat{\theta}_{OLS}$ and $\hat{\sigma}^2 = \frac{\ell^T\ell}{n-p}$ which is the MSE of OLS regression model.
Several influential diagnostics have been proposed based on the OLS regression to detect the impact of deletion on the regression analysis but some of the Two-Parameter estimators have not been considered of which New Two-Parameter (NTP) estimator is among. As a result of this, generalized versions of some diagnostic influential measures of DFFITs (DFT) and Cook's D (CKD), COVRATIO, Hadi's measure (HAD), Pena's Statistic (PEN), and Atkinson statistic (ATK) with the New Two-Parameter estimator (NTP) are hereby proposed in this study.

The fitted value of the equation (16) can expressed as follows:
$\hat{Y}^{NTP}_i = W\hat{\theta}_{NTP}$,

$$= W\left(W^TW + I\right)^{-1}\left(W^TW + dI\right)\left(W^TW + kI\right)^{-1}W^TY,$$

$$\hat{Y}_i^{NTP} = H_{ii}^{NTP} Y$$

(21)

where,

$$H_{ii}^{NTP} = W\left(W^T W + I\right)^{-1}\left(W^T W + dI\right)\left(W^T W + kI\right)^{-1} W^T$$

.

H_{ii}^{NTP} is equivalent to the hat matrix $W\left(W^T W\right)^{-1} W^T$ of the fitted value of OLS when *k = 0* and *d= 1*. It is expected to note that H_{ii}^{NTP} is not an idempotent matrix, hence it is not a projection matrix. The *i^th^* diagonal element of H_{ii}^{NTP} can be written as:

$$h_{ii}^{NTP} = w_i \left(W^T W + I\right)^{-1}\left(W^T W + dI\right)\left(W^T W + kI\right)^{-1} w^T$$

(22)

where w_i indicates the *i*th row of the matrix *W*, the ith fitted value written in terms of H_{ii}^{NTP} is given as:

$$\hat{Y}_i^{NTP} = \sum_{j=1}^{n} h_{ii}^{NTP} Y_i \, .$$

(23)

Also, minimizing (23) with respect to Y_i , leads to (24):

$$\frac{\partial \hat{Y}_i^{NTP}}{\partial Y_i} = h_{ii}^{NTP} \, ,$$

(24)

such that the ith NTP error term is given as:

$$\ell_i^{NTP} = Y_i - Y_i^{NTP} \, ,$$

(25)

$$\ell_i^{NTP} = (1 - h_{ii}^{NTP}) Y_i$$

(26)

3.1 DFFITs and Cook's distance measures in NTP Estimator

Following the conventional DFFITs measure developed based on OLS, therefore DFFITs for NTP estimator denoted as DFTNTP is hereby proposed and as expressed as:

$$DFFITs_{iNTP} = DFTNTP_i = \frac{w_i\left(\hat{\theta}_{NTP} - \hat{\theta}_{(i)NTP}\right)}{SE\left(w_i \hat{\theta}_{NTP}\right)}$$

(27)

where

$$SE\left(w_i \hat{\theta}_{NTP}\right) = s\sqrt{\left(w_i\left(W^T W + I\right)^{-1} E_0 \left(W^T W\right)^{-1} E_1 \left(W^T W + I\right)^{-1} w^T\right)}$$

such that, $E_0 = \left(W^T W + dI\right)\left(W^T W + kI\right)^{-1}$ and $E_1 = \left(W^T W + kI\right)^{-1}\left(W^T W + dI\right)$. This is an estimator of the standard error of the NTP estimator fitted values. It can still be written as:

$$SE\left(w_i \hat{\theta}_{NTP}\right) = s_{(i)} \sqrt{\sum_{j=1}^{n} \left(h_{ij}^{NTP}\right)^2} \, ,$$

(28)

where $s_{(i)} = \sqrt{\dfrac{(n-p)s^2 - \ell_i^2 / (1 - h_{ii})}{n-p-1}}$, such that h_{ij}^{NTP} is the *ij^th^* element of H_{ii}^{NTP} and *s* is the unbiased OLS estimator of σ.

Also, further algebraic expression of (27) is given in (28).

$$DFFITs_i^{NTP} = \frac{\hat{y}_i^{NTP} - \hat{y}_{(i)}^{NTP}}{s_{(i)NTP}\sqrt{h_{ii}^{NTP}}}$$

(29)

where $\hat{y}_i^{NTP}$ and $\hat{y}_{(-i)}^{NTP}$ are the NTP estimator predicted values of response variable *y* when no observation is omitted and when the *i*th observation ,$s_{(-i)NTP}$ is the NTP estimator standard error estimated when the ith data is deleted and h_{ii}^{NTP} is the *i*th diagonal elements of the hat matrix H_{ii}^{NTP} .

$|DFFITs_i^{NTP}| > 2\sqrt{\frac{p}{n}}$ is estimated to be the cut-off point for large data.

In the same vein, the Cook's distance version for NTP estimator can be expressed as follows:

$$D_i^{NTP} = CKDNTP_i = \frac{\left(\hat{\theta}_{NTP} - \hat{\theta}_{(i)NTP}\right)^T \left(W^T W\right)\left(\hat{\theta}_{NTP} - \hat{\theta}_{(i)NTP}\right)}{ps^2}$$

(30)

Likewise, the other version of D_i^{NTP} is defined as:

$$D_i^{*NTP} = \frac{\left(\hat{\theta}_{NTP} - \hat{\theta}_{(i)NTP}\right)^T \left(W^T W + I\right) E_0 \left(W^T W\right) E_1 \left(W^T W + I\right)^{-1}\left(\hat{\theta}_{NTP} - \hat{\theta}_{(i)NTP}\right)}{ps^2}$$

(31a)

$$CKDNTP^{*NTP} = \frac{\left(\hat{\theta}_{NTP} - \hat{\theta}_{(i)NTP}\right)^T \left(W^T W + I\right) E_0 \left(W^T W\right) E_1 \left(W^T W + I\right)^{-1}\left(\hat{\theta}_{NTP} - \hat{\theta}_{(i)NTP}\right)}{ps^2}$$

(31b)

where $D_i^{*NTP} = CKDNTP^{*NTP}$

In this case, D_i^{*NTP} serves as an alternative influential measure to Cook's distance based on;

$$Var\left(\hat{\theta}_{NTP}\right)=\sigma\left(W^TW+I\right)^{-1}E_0\left(W^TW\right)^{-1}E_1\left(W^TW+I\right)^{-1}$$

. The measures D_i^{NTP} and D_i^{*NTP} cannot be written as functions of leverage and residual, this is because of scale dependency of NTP estimator. The estimator is not scale invariant. Hence, the design matrix *W* with *i*th row deleted needs to be rescaled before $\hat{\theta}_{(i)NTP}$ is computed.

The cut-off point is $D_i^{NTP} > 1$ this implies that the observation that its $D_i^{NTP} > 1$ when sample size (n) is large is influential and thereby warrants attention.

The approximate case deletion formula for NTP estimator is hereby presented:

Approximate Case Deletion formulas for NTP estimator

Given that *i*th row is deleted from $\hat{\theta}_{NTP}$, $\hat{\theta}_{(i)NTP}$ can then be written as:

$$\hat{\theta}_{(i)NTP}=\left(W_{(i)}^TW_{(i)}+I\right)^{-1}\left(W_{(i)}^TW_{(i)}+dI\right)\left(W_{(i)}^TW_{(i)}+kI\right)^{-1}W_{(i)}^TY_{(i)}$$

(32)

where and $W_{(i)}$ is the Matrix W when the ith row has been removed. W is scaled so that $W_{(i)}W$ is in correlation form.

Therefore, by applying Sherman-Morrison-Woodbury (SMW) theorem by [47], $\hat{\theta}_{(i)NTP}$ is approximately obtained as follows:

$$\hat{\theta}_{(i)NTP}=\left(W^TW-w_i^Tw_i+I\right)^{-1}\left(W^TW-w_i^Tw_i+dI\right)\left(W^TW+kI\right)^{-1}W^TY$$

(33)

$$\hat{\theta}_{(i)NTP}=\left(B_k-w_i^Tw_i\right)^{-1}\left(W^TW-w_i^Tw_i+dI\right)\left(W^TW+kI\right)^{-1}W^TY$$

(34)

where, $B_k=\left(W^TW+I\right)$,

$$\hat{\theta}_{(i)NTP}=\left(B_k^{-1}+\frac{B_k^{-1}w_i^Tw_iB_k^{-1}}{1-w_iB_k^{-1}w_i^T}\right)\left(W^TW-w_i^Tw_i+dI\right)\left(W^TW+kI\right)^{-1}W^TY$$

(35)

where, $\left(B_k-w_i^Tw_i\right)^{-1}=B_k^{-1}+\dfrac{B_k^{-1}w_i^Tw_iB_k^{-1}}{1-w_iB_k^{-1}w_i^T}$,

$$\hat{\theta}_{(i)NTP}=\left(B_k^{-1}+\frac{B_k^{-1}w_i^Tw_iB_k^{-1}}{1-F_{ii}}\right)\left(W^TW-w_i^Tw_i+dI\right)\left(W^TW+kI\right)^{-1}W^TY$$

(36)

where, $F_{ii}=w_iB_k^{-1}w_i^T$,

By expanding (36), it leads to (37).

$$\hat{\theta}_{(i)NTP}=\left(B_k^{-1}W^TW-w_i^Tw_iB_k^{-1}+dIB_k^{-1}+\frac{B_k^{-1}w_i^Tw_iB_k^{-1}}{1-F_{ii}}\left(W^TW\right)-\frac{w_i^Tw_iB_k^{-1}w_i^Tw_iB_k^{-1}}{1-F_{ii}}+\frac{dIB_k^{-1}w_i^Tw_iB_k^{-1}}{1-F_{ii}}\right)\hat{\theta}_{ORR}$$

(37)

where, $\hat{\theta}_{ORR}=\left(W^TW+kI\right)^{-1}W^TY$,

$$\hat{\theta}_{(i)NTP}=\left[B_k^{-1}\left(W^TW+dI\right)+\frac{B_k^{-1}w_i^T}{1-F_{ii}}\begin{pmatrix}dIB_k^{-1}w_i+w_iB_k^{-1}\left(W^TW\right)\\-w_iB_k^{-1}w_i^Tw_i-\left(1-F_{ii}\right)w_i\end{pmatrix}\right]\hat{\theta}_{ORR}$$

(38)

$$=B_k^{-1}\left(W^TW+dI\right)\hat{\theta}_{ORR}+$$

$$\frac{B_k^{-1}w_i^T}{1-F_{ii}}\begin{bmatrix}dIB_k^{-1}w_i+w_iB_k^{-1}\left(W^TW\right)-\\w_iB_k^{-1}w_i^Tw_i\left(1-F_{ii}\right)w_i\end{bmatrix}\hat{\theta}_{ORR},$$

(39)

$$=$$

$$\hat{\theta}_{NTP}+\frac{B_k^{-1}w_i^T}{1-F_{ii}}\left[w_iB_k^{-1}\left(W^TW+dI\right)-w_iB_k^{-1}w_i^Tw_i-\left(1-F_{ii}\right)w_i\right]\hat{\theta}_{ORR}$$

(40)

where,

$$\hat{\theta}_{NTP}=B_k^{-1}\left(W^TW+dI\right)\hat{\theta}_{ORR}=\left(W^TW+I\right)^{-1}\left(W^TW+dI\right)\left(W^TW+kI\right)^{-1}W^TY$$

,

$$\hat{\theta}_{(i)NTP}=\hat{\theta}_{NTP}-\frac{B_k^{-1}w_i^T}{1-F_{ii}}\left[w_iB_k^{-1}\left(W^TW+dI\right)-w_iB_k^{-1}w_i^Tw_i-\left(1-F_{ii}\right)w_i\right]\hat{\theta}_{ORR}$$

(41)

$$=\hat{\theta}_{NTP}-\frac{B_k^{-1}w_i^T}{1-F_{ii}}\left[-w_iB_k^{-1}\left(W^TW+dI\right)+w_iF_{ii}+\left(1-F_{ii}\right)w_i\right]\hat{\theta}_{ORR}$$

$$=\hat{\theta}_{NTP}-\frac{B_k^{-1}w_i^T}{1-F_{ii}}\left[-w_iB_k^{-1}\left(W^TW+dI\right)+w_i\right]\hat{\theta}_{ORR}$$

$$=\frac{B_k^{-1}w_i^T}{1-F_{ii}}\left[w_i-w_iB_k^{-1}\left(W^TW+dI\right)\right]\hat{\theta}_{ORR}$$

(42)

$$\hat{\theta}_{NTP}-\hat{\theta}_{(i)NTP}=\frac{B_k^{-1}w_i^T}{1-F_{ii}}\left[w_i-w_iB_k^{-1}\left(W^TW+dI\right)\right]\hat{\theta}_{ORR}$$

(43)

$$=\frac{B_k^{-1}w_i^T}{1-F_{ii}}\left[w_i\hat{\theta}_{ORR}-w_iB_k^{-1}\left(W^TW+dI\right)\hat{\theta}_{ORR}\right]$$

(44)

Therefore, the approximate difference between $\hat{\theta}_{NTP}$ and $\hat{\theta}_{(i)NTP}$ is given as:

$$\hat{\theta}_{NTP}-\hat{\theta}_{(i)NTP}\cong\frac{B_k^{-1}w_i}{1-F_{ii}}\left(1-B_k^{-1}\left(W^TW+dI\right)\right)w_i^T\hat{\theta}_{ORR}$$

(45)

where, $B_k^{-1} = \left(W^T W + I\right)^{-1}$, $\hat{\theta}_{ORR} = \left(W^T W + kI\right)^{-1} W^T Y$ and $F_{ii} = w_i \left(W^T W + I\right)^{-1} w_i^T$.

Approximate case deletion formula for DFFITs in NTP Estimator

The approximate case deletion formula for DFFITs in NTP estimator can be expressed as:

$$DFFITs_{iNTP}^{apm} = \left\lfloor \frac{F_{ii}}{1 - F_{ii}} \right\rfloor \frac{\ell_i^{NTP}}{s\left(w_i\, \hat{\theta}_{NTP}\right)},$$

$$= \left\lfloor \frac{F_{ii}}{1 - F_{ii}} \right\rfloor \frac{\ell_i^{NTP}}{s_{(i)} \sqrt{\sum_{j=1}^{n} h_{ij}^{NTP}}}$$

(46)

where, $s_{(i)} = \sqrt{\frac{(n-p)s^2 - \frac{\ell_i}{1 - h_{ii}}}{n - p - 1}}$,

$F_{ii} = w_i \left(W^T W + I\right)^{-1} w_i^T$ and $\ell_i^{NTP} = Y_{(i)} - \hat{Y}^{NTP}$

Approximate case deletion formula for Cook's distance in NTP Estimator

The approximate version for Cook's distance in NTP estimator is hereby presented as in (47)

$$D_{iNTP}^{apm} = \frac{1}{ps^2} \left\lfloor \left[\frac{B_k^{-1} w_i^T}{1 - F_{ii}} \left[w_i \left(1 - B_k^{-1} \left(W^T W + dI\right)\right) \hat{\theta}_{ORR} \right] \right]^T \left(W^T W\right) \left[\frac{B_k^{-1} w_i^T}{1 - F_{ii}} \left[w_i \left(1 - B_k^{-1} \left(W^T W + dI\right)\right) \hat{\theta}_{ORR} \right] \right] \right\rfloor \quad (47)$$

Therefore, the approximate version of (47) can be written as:

$$D_{iNTP}^{apm} = \frac{1}{ps^2} \left[\frac{\ell_i^{NTP}}{1 - F_{ii}} \right]^2 \sum_{j=1}^{n} \left(h_{ij}^{NTP}\right)^2$$

(48)

where, $\ell_i^{NTP} = (1 - h_{ii}^{NTP}) Y_i$

$$h_{ii}^{NTP} = w_i \left(W^T W + I\right)^{-1} \left(W^T W + dI\right) \left(W^T W + kI\right)^{-1} w^T$$

and $F_{ii} = w_i \left(W^T W + I\right)^{-1} w_i^T$.

3.2 COVRATIO measure in NTP Estimator

The generalized version of COVRATIO diagnostic measure in NTP estimator can be expressed as in (49). This is achieved by following the procedures of conventional COVRATIO measures proposed based on OLS by [14]. The influential measure is denoted as CVRNTP and defined as:

$$COVRATIO_i^{NTP} = CVRNTP_i = \frac{(S_{(-i)NTP}^2)^{(p+1)}}{MS\hat{E}_{NTP}^{(p+1)}} \left(\frac{1}{1 - h_{ii}^{NTP}} \right)$$

(49)

where $MS\hat{E}_{NTP}^{(p+1)} = \frac{SSE_{NTP}}{n - p + 1}$,

such that $SSE_{NTP} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i^{NTP})^2$ and $S_{(-i)NTP}^2$ is the variance for the NTP estimator when the i^{th} observation is deleted.

Therefore, observation for which $\left| COVRATIO_i^{NTP} - 1 \right| \geq \frac{3p}{n}$ is said to be influential and this is the cut point for the influential measure.

3.3 Hadi's measure in NTP Estimator

In line with the conventional Hadi's measure proposed based on OLS by [23], therefore the generalized version of Hadi's measure in NTP estimator is hereby proposed which is denoted as HADNTP and expressed as:

$$H_{iNTP}^2 = HADNTP_i^2 = \frac{h_{ii}^{NTP}}{1 - h_{ii}^{NTP}} + \left(\frac{p}{1 - h_{ii}^{NTP}} \right) \left(\frac{d_{iNTP}^2}{1 - d_{iNTP}^2} \right)$$

(50)

Where $H_{iNTP}^2 = HADNTP_i^2$

Such that $d_{iNTP} = \frac{e_{iNTP}}{\sqrt{SSE}_{iNTP}}$, such that $e_{iNTP} = Y_i - \hat{Y}_i^{NTP}$, $SSE_{iNTP} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^{NTP}\right)^2$ and p is the number of explanatory variable and h_{ii}^{NTP} is the hat matrix for NTP estimator.

The point at which $H_{iNTP}^2 = mean(H_{iNTP}^2) + c\sqrt{\operatorname{var}(H_{iNTP}^2)}$ is influential, where c is an appropriately chosen constant to be 2.

3.4 Pena's Statistic version in NTP estimator

Following the procedures of [23] and [22], The generalized version of Pena's Statistic in NTP estimator is hereby expressed as:

$$S_{iNTP} = PEN_{iNTP} = \frac{1}{p\hat{\sigma}^2 h_{ii}^{NTP}} \sum_{j=1}^{n} \frac{(h_{ji}^{NTP})^2 e_{jNTP}^2}{(1-h_{ji}^{NTP})^2} .$$

(51)

where $\ell_{jiNTP} = Y_i - \hat{Y}_i^{NTP}$

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2}{n-p}$, h_{ii}^{NTP} is the ith diagonal element of the hat matrix H_{ii}^{NTP} .and h_{ji}^{NTP} is the jth element of the hat matrix H_{ii}^{NTP}. The cutoff point for the influential measure is given as;

$$\left|S_{iNTP} - median(S_{iNTP})\right| \geq 4.5MAD(S_{iNTP}) .$$

3.5 Atkinson Diagnostic (ATK) in NTP Estimator

The generalized version of Atkinson diagnostic (ATK) for the NTP estimator is hereby proposed by following the conventional Atkinson diagnostic proposed with the OLS estimator by [7]. The influential measure is presented as:

$$ATK_i^{NTP} = ATKNTP_i = \frac{t^2_{(-i)NTP}}{p}\left(\frac{h_{ii}^{NTP}}{1-h_{ii}^{NTP}}\right).$$

(53)

such that $t_{(-i)NTP} = \frac{\ell^{NTP}_{(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)NTP}\left(1-\mathrm{h}_{\mathrm{ii}}^{\mathrm{NTP}}\right)}}$,

$e^{NTP}_{(-i)} = Y_i - \hat{Y}^{NTP}_{(-i)}$ and $\hat{\sigma}^2_{(-i)NTP} = \frac{(\ell^{NTP}_{(-i)})^T(\ell^{NTP}_{(-i)})}{n-p}$

The observation that its ATK_i^{NTP}>1 identifies to be influential, this is the cut point for the influential measure.

3.6 Simulation Procedures

This study's simulation approach was implemented using the R statistical programming language. Equation (53) was used to produce all of the exogenous variables as also done by [39].

$$w_{ij} = \sqrt{(1-r^2)}z_{ij} + rz_{i(j+1)}, i = 1, 2, ..., n, j = 1, 2,..., p.$$

(54)

ρ indicates the correlation between any two exogenous variables, and z_{ij} represents independent standard normal pseudo-random integers in this equation. To demonstrate the degrees of correlations between the regressors, five levels of correlations (r=0,0.8,0.9,0.95,0.99) were used, and p = 3 denotes the number of numbers of regressors. The variables had a standard form of expression. In a similar manner, the response variable was produced using the following equation:

$$y = \theta_0 + \theta_1 w_1 + \theta_2 w_2 + \theta_3 w_3 + ... + \theta_p w_{ip} + e_i , i = 1,2,3,..$$

(55)

ℓ_i is the residual, which is independently and identically normally distributed with mean (0) and variance σ^2 that is $\ell_{i_i} \sim i\,idN(0,\hat{\sigma}^2)$. For the model in (55), zero intercept was used, and values of β were selected to satisfy the criteria $Q^TQ = 1$ in order to comply with [39] guidelines. The simulation experiments were repeated 1000 times for the sample sizes n = 10, 20, 30, 40, 50, 100, 250, and 500, respectively, with a standard deviation of 1, 5, and 10 pertaining to the research conducted by [30], [39], [11], [15] and others. In order to include outliers in the regressors, equation (56) was employed which has been used by several authors such as [9], [3] among all other authors.

$W(i)_{outlier}$ = mg*Max (W_i) + W_i

(56)

where mg, which is assigned the numbers 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 which represents the magnitude of outliers in the x-direction. Whereas gx = 10% and 20% means the percentage of outliers that is 10% and 20% of the data generated were randomly selected and polluted with outliers using equation (56). Different diagnostic tools considered in the study were acquired and subjected to their various cut-off points. Similarly, the number of influential points found was divided by the number of inflated outliers to get the percentage (%) of influential points. The best influential measure was determined using the highest count at 100% influential point detection.

3.6.1 Algorithm for the Generation of Explanatory Variables, Response Variables, Error Terms, and Mean Squares Error

(i) Choose the sample size to work with, say n.

(ii) Generate exogenous variables using equation

$$w_{ij} = (1-r^2)^{\frac{1}{2}} z_{ij} + rz_{ip+1},$$

(iii) Choose a percentage of the data to be replaced with outliers.

(iv) Choose a particular magnitude of outlier to invoke into the data to be randomly generated from step 2

(v) Randomly select those observations making up the percentage of the generated data to be replaced with outliers.

(vi) Invoke outliers into the data using;
$W(i)_{\text{outlier}} = mg*\text{Max}\,(W_i) + W_i$
where mg is the magnitude of outliers in the x-direction

(vii) Replace the outliers in the original data generated in (iii)

(viii) Generate response variable using:

$$y = \theta_0 + \theta_1 w_1 + \theta_2 w_2 + \theta_3 w_3 + ... + \theta_p w_{ip} + e_i \,, i = 1, 2, 3, ..., p.$$

(ix) Obtain the MSE of the estimators of the model in step (viii)

(x) Compute for each replicate the estimated MSE for each of the estimators by dividing the result in step (ix) by the number of replications.

(xi) Choose another magnitude of outliers to invoke into the data and repeat step (iv) to step (x).

(xii) Repeat step (iv) to (xi) until all the magnitudes of outliers are exhausted.

(xiii) Repeat steps (ii) to (xii) until all the sample sizes are exhausted.

4. Results and Discussion

Tables 1 – 6 show the samples of simulation results of the percentage of influential points detected by some already existing diagnostic tools proposed based on OLS and the newly proposed ones based on NTP. From the tables, it is evident that certain influential diagnostic measures, such as CKD, PEN, and ATK (proposed based on OLS), effectively detect influential points only when the sample size is small (e.g., 10 or 20) and in the absence of multicollinearity and outliers. However, their performance declines when multicollinearity is controlled, as reflected in Table 7. Analyzing the detection percentages in Table 7 reveals variability in the effectiveness of the measures. CKD, PEN, ATK, and ATKNTP demonstrated inconsistent performance, performing weakly except within the 0–9.99% detection range, where their total counts were 2495, 2499, 2300, 2515, and 2501, respectively (overall total expected =levels of multicollinearity x levels of error variance x levels of outliers x levels of magnitude of outliers x levels of sample size=5x3x11x2x8=2640). This indicates a limited capability for detecting influential points, likely influenced by factors such as error variance, multicollinearity, outliers, and sample size. In contrast, DFT, DFTNTP, HAD, and HAD performed well across various detection categories. Additionally, CVR and CVRNTP showed strong performance, particularly at 90–99.99% and 100% detection levels. Notably, CVRNTP outperformed the others at 100% detection, achieving the highest total count of 1970. This highlights CVRNTP's superior ability to detect influential points when multicollinearity is mitigated, aligning with the findings of [58]. In the same vein, Table 8 reveals the summary of the performance of the diagnostic measures that have 100% influential points detection when counted over all the specifications such as percentage of outliers, magnitudes of outliers, correlation levels, and error variances at each sample size. Hence, it can be seen that the proposed CVRNTP has the highest counts among others, with a wide margin of the total count of 1970 followed by CVR. Also, Figure 1 is the bar chart of the summary of results.

Table 1: Percentage of influential points detected when *n*=10, *r*=0, σ = 5

		OLS						NTP					
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
0.1	0	2.9	32.9	0	80.7	0	4	0.2	1.2	0.8	98.8	0	0
	1	0	9.3	0	73	0	0.2	0	0.3	2.8	97.7	0	0
	2	0	20.2	0	79.2	0	1.1	0.1	1.9	19.1	98.9	0	0
	3	15.3	46.9	100	87.7	0	12.3	1.4	7.7	78	99.2	0	0.5
	4	38.4	63.1	100	92.5	100	45.4	3.6	15.9	82.8	99.7	2.1	1
	5	52.5	70.3	100	94.5	100	70	5.4	24.5	87.8	99.8	5.3	3.3
	6	61.2	75.9	100	96.2	100	83	8.8	35	92.7	99.9	9.7	5.7
	7	67.5	79.4	100	97	100	90	13.1	46.1	95.8	100	14.2	8.7
	8	71.9	82.4	100	97.6	100	93.7	18	56.9	96.1	100	18.3	13.3
	9	74.6	85	100	98	100	95.9	24.3	65.1	98.8	100	24.1	18.7
	10	76.9	86.5	100	98.2	100	96.9	30.3	70.9	99.4	100	28.4	25.4

		OLS						NTP					
	mg	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**

0.2	0	0	10.4	0	74.4	0	0.65	0	0.4	2.1	97.85	0	0.05
	1	1.8	22.45	0	78.9	0	2.35	0.2	2.85	34.5	98.45	0	0.05
	2	3.5	31.05	41.6	82.15	0	4.1	0.2	7.05	42.2	98.95	0	0.05
	3	3.6	33.4	0.85	82.85	0	4.1	0.55	10.95	45.45	98.6	0	0.1
	4	4	34.25	0.05	83.45	0	3.95	0.95	14.1	47.4	98.7	0	0.15
	5	4.2	34.95	0	83.75	0	4.05	1.35	16.55	48.55	98.7	0	0.15
	6	4.15	35.2	0	84	0	4.15	1.8	17.75	48.5	98.6	0	0.15
	7	4.45	35.7	0	83.9	0	4.15	2.15	18.6	48.55	98.65	0	0.25
	8	4.65	35.95	0	84.25	0	4.1	2.45	20.1	48.5	98.7	0	0.3
	9	4.65	35.95	0	84.1	0	4.1	2.65	20.5	48.1	98.8	0	0.35
	10	4.65	36.05	0	84.1	0	4	2.8	21.1	47.9	98.75	0	0.35

Table 2: Percentage of influential points detected when *n*=20, *r*=0, σ = 1

			OLS							NTP			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
0.1	0	0	2.05	0	96.75	0	0.15	0	0.55	0.8	99.85	0	0
	1	0	15.25	0	97.15	0	3.45	0	6.65	3.2	99.75	0	0.3
	2	0.7	30.5	50	98.15	0	16.2	0.35	43.85	55.45	99.95	0	2.5
	3	1.85	37.3	50	98.65	0	23.55	2.3	71.5	72.8	99.9	0	5
	4	3	40	50	98.75	0	27	7.85	77.75	89.35	99.9	0	8.1
	5	3.1	41.55	50.2	98.7	0	29	16.35	79.55	97.25	99.95	0	10.8
	6	3.55	42.65	52.2	98.7	0	30.5	27.35	79.6	99.45	99.95	0	13.7
	7	3.75	43.1	65.1	98.75	0	31.3	36.15	78.85	99.85	99.95	0	15.5
	8	3.85	42.9	96.3	98.75	0	31.9	40.55	77.85	99.9	99.95	0	17.3
	9	3.85	43	99.7	98.8	0	32.45	42.75	76.75	99.95	99.95	0	19.1
	10	3.9	42.95	100	98.85	0	33.25	43.1	75.4	100	99.95	0	20.1
			OLS							**NTP**			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
0.2	0	2.65	21.38	25	97.15	0	13.65	0.025	4.125	8.75	99.975	0	0.2
	1	3.05	29.08	25	97.6	0	18.3	0.225	17.025	25.13	99.7	0	3.03
	2	3.65	33.9	25	97.95	0	23.225	0.175	36.475	32.4	99.975	0	1.25
	3	4.15	36.55	25.5	98.25	0	25.4	0.3	47.9	36.65	99.95	0	1.53
	4	4.25	37.68	27	98.35	0	26.75	0.525	53.775	38.2	99.925	0	1.93
	5	4.35	38.38	29.2	98.375	0	27.3	1.2	56.425	38.6	99.9	0	2.25
	6	4.4	38.68	31.4	98.475	0	27.525	2.625	56.825	38.7	99.85	0	2.25
	7	4.43	39.08	33	98.525	0	27.925	5.575	56.7	38.55	99.8	0	2.38
	8	4.43	39.28	34.3	98.55	0	28.075	8.55	55.975	38.18	99.8	0	2.45
	9	4.43	39.45	34.8	98.575	0	28.25	11.23	55.05	37.03	99.8	0	2.53
	10	4.38	39.43	34.9	98.625	0	28.25	14	53.925	36	99.8	0	2.53

Table 3: Percentage of influential point detected when n=40, *r* =0.8, σ = 5

			OLS							NTP			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
0.1	0	0	15.93	0	100	0	0	0	13.625	1.4	100	0	0
	1	0	28.03	25	100	0	0	0	15.525	21.28	100	0	0
	2	0.33	39.48	75	100	0	0	0	24.975	24.33	100	0	0
	3	0.35	42.23	75	100	0	0	0	27.65	66.5	100	0	0
	4	0.38	43.28	75	100	0	0	0.025	31.625	91.7	100	0	0

	5	0.38	43.83	75	100	0	0	0.025	35.75	95.35	100	0	0
	6	0.33	44.15	75	100	0	0	0.025	39.4	96.83	100	0	0
	7	0.33	44.5	75	100	0	0	0.025	42.35	97.68	100	0	0
	8	0.33	44.6	75	100	0	0	0.025	43.85	98.23	100	0	0
	9	0.33	44.68	75	100	0	0	0.025	45.15	98.73	100	0	0
	10	0.33	44.85	75	100	0	0	0.025	46.225	99.08	100	0	0

			OLS							NTP			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
	0	0	8.038	0	100	0	0	0	12.525	9.95	100	0	0
	1	0	18.46	12.5	99.988	0	0	0	17.763	12.61	100	0	0
	2	0	20.93	12.5	99.988	0	0	0	16.75	36.68	100	0	0
0.2	3	0	21.74	12.5	99.988	0	0	0	15.6	40.38	100	0	0
	4	0	21.98	12.5	99.988	0	0	0	15.825	40.9	100	0	0
	5	0.01	22.03	12.9	99.988	0	0	0	16.488	40.88	100	0	0
	6	0.01	22.13	25	99.988	0	0	0	17.488	40.63	100	0	0
	7	0.01	22.16	25	99.988	0	0	0	18.325	40.3	100	0	0
	8	0.01	22.14	25	99.988	0	0	0	19.188	39.88	100	0	0
	9	0.01	22.15	25	99.988	0	0	0	19.963	39.54	100	0	0
	10	0.01	22.18	25	99.988	0	0	0	20.413	38.94	100	0	0

Table 4: Percentage of influential point detected when n=100, r =0.9, σ = 10

			OLS							NTP			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
	0	0	10	0	100	0	0	0	0	10	100	0	0
	1	0	20	50	100	0	0	0	20	20	100	0	0
	2	0	30	60	100	0	0	0	20	50	100	0	0
0.1	3	0	30	60	100	0	0	0	20	80	100	0	0
	4	0	30	80	100	0	0	0	20	80	100	0	0
	5	0	30	80	100	0	0	0	20	100	100	0	0
	6	0	30	80	100	0	0	0	30	100	100	0	0
	7	0	30	80	100	0	0	0	30	100	100	0	0
	8	0	30	80	100	0	0	0	30	100	100	0	0
	9	0	30	80	100	0	0	0	30	100	100	0	0
	10	0	30	80	100	0	0	0	30	100	100	0	0

			OLS							NTP			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
	0	0	10	0	100	0	0	0	0	0	100	0	0
	1	0	20	15	100	0	0	0	5	10	100	0	0
	2	0	20	10	100	0	0	0	5	30	100	0	0
	3	0	20	10	100	0	0	0	5	35	100	0	0
0.2	4	0	20	10	100	0	0	0	5	35	100	0	0
	5	0	20	10	100	0	0	0	10	35	100	0	0

	6	0	20	10	100	0	0	0	10	35	100	0	0
	7	0	20	10	100	0	0	0	10	40	100	0	0
	8	0	20	10	100	0	0	0	10	40	100	0	0
	9	0	20	10	100	0	0	0	5	40	100	0	0
	10	0	20	10	100	0	0	0	5	40	100	0	0

Table 5: Percentage of influential point detected when n=250, r =0.95, σ = 5

			OLS							NTP			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
0.1	0	0	10.14	8	100	0	0	0	29.04	4.14	100	0	0
	1	0	31.28	60	100	0	0	0	45.812	27.02	100	0	0
	2	0	32.59	60	100	0	0	0	43.784	79.82	100	0	0
	3	0	32.8	56	100	0	0	0	37.856	93.42	100	0	0
	4	0	32.84	52	100	0	0	0	36.452	97.55	100	0	0
	5	0	32.9	52	100	0	0	0	37.52	98.36	100	0	0
	6	0	32.92	52	100	0	0	0	38.192	99.33	100	0	0
	7	0	32.94	52	100	0	0	0	38.944	99.54	100	0	0
	8	0	33.02	52	100	0	0	0	39.308	99.7	100	0	0
	9	0	33.01	48	100	0	0	0	39.316	99.8	100	0	0
	10	0	32.99	48	100	0	0	0	39.108	99.83	100	0	0
			OLS							**NTP**			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
0.2	0	0	10.35	4	100	0	0	0	31.312	2.234	100	0	0
	1	0	19.61	20	100	0	0	0	43.796	23.63	100	0	0
	2	0	19.89	16	100	0	0	0	29.372	29.77	100	0	0
	3	0	19.93	16	100	0	0	0	22.288	30.51	100	0	0
	4	0	19.9	16	100	0	0	0	20.51	30.7	100	0	0
	5	0	19.9	16	100	0	0	0	20.304	30.86	100	0	0
	6	0	19.92	16	100	0	0	0	20.476	30.75	100	0	0
	7	0	19.93	16	100	0	0	0	20.618	30.6	100	0	0
	8	0	19.93	16	100	0	0	0	20.68	30.56	100	0	0
	9	0	19.93	16	100	0	0	0	20.736	30.46	100	0	0
	10	0	19.93	16	100	0	0	0	20.772	30.38	100	0	0

Table 6: Percentage of influential point detected when n=500, r =0.99, σ = 10

			OLS							NTP			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
0.1	0	0	8.906	6	100	0	0	0	10.708	2.232	100	0	0
	1	0	31.41	64	100	0	0	0	40.664	25.86	100	0	0
	2	0	31.64	66	100	0	0	0	43.006	89.05	100	0	0
	3	0	31.68	58	100	0	0	0	39.928	98.18	100	0	0
	4	0	31.69	56	100	0	0	0	36.944	99.01	100	0	0
	5	0	31.7	58	100	0	0	0	35.298	99.56	100	0	0
	6	0	31.71	60	100	0	0	0	34.274	99.62	100	0	0
	7	0	31.7	60	100	0	0	0	34.028	99.78	100	0	0

	8	0	31.71	60	100	0	0	0	33.696	99.69	100	0	0
	9	0	31.71	60	100	0	0	0	33.558	99.77	100	0	0
	10	0	31.72	60	100	0	0	0	33.608	99.93	100	0	0
			OLS							**NTP**			
gx	**mg**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
	0	0	10.09	9	100	0	0	0	13.257	10.32	100	0	0
	1	0	19.7	16	100	0	0	0	34.809	19.08	100	0	0
	2	0	19.8	15	100	0	0	0	31.797	23.91	100	0	0
0.2	3	0	19.79	15	100	0	0	0	26.79	24.93	100	0	0
	4	0	19.81	15	100	0	0	0	23.191	25.67	100	0	0
	5	0	19.81	15	100	0	0	0	20.794	26.19	100	0	0
	6	0	19.81	15	100	0	0	0	19.588	26.24	100	0	0
	7	0	19.82	15	100	0	0	0	18.842	26.28	100	0	0
	8	0	19.82	15	100	0	0	0	18.445	26.26	100	0	0
	9	0	19.81	15	100	0	0	0	18.157	26.14	100	0	0
	10	0	19.82	15	100	0	0	0	18.012	26.05	100	0	0

Table 7: Percentage of influential points detected when counted overall specifications

	OLS						**NTP**					
% of detection	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**	**CKD**	**DFT**	**HAD**	**CVR**	**PEN**	**ATK**
100	0	0	377	1573	134	29	0	17	103	1970	0	0
90-99.99	39	60	40	870	0	82	0	36	547	668	0	0
80-89.99	36	48	48	168	0	11	5	15	113	0	17	3
70-79.99	30	21	180	28	0	7	8	74	49	1	10	14
60-69.99	15	9	188	0	0	4	9	109	314	0	24	9
50-59.99	9	3	203	0	0	17	5	134	58	0	25	4
40-49.99	3	438	162	0	0	36	4	234	267	0	18	4
30-39.99	6	852	81	1	0	44	10	556	349	1	14	7
20-29.99	3	709	247	0	3	60	26	475	415	0	8	13
10-19.99	4	451	487	0	4	50	78	638	162	0	9	85
0-9.99	2495	49	627	0	2499	2300	2495	352	263	0	2515	2501
Total	**2640**	**2640**	**2640**	**2640**	**2640**	**2640**	**2640**	**2640**	**2640**	**2640**	**2640**	**2640**

Table 8: Frequency of the influential measures that correctly detected 100% influential point when counted over all (mg), (gx), (r), and (σ) at each sample sizes (n)

	Sample size (n)									
Influential Measures	**10**	**20**	**30**	**40**	**50**	**100**	**250**	**500**	**Total**	**Rank**
CVRNTP	96	0	226	330	330	330	329	329	1970	1
CVR	9	0	9	237	330	330	329	329	1573	2
HADNTP	0	1	3	3	1	86	9	0	103	5
HAD	126	94	110	0	47	0	0	0	377	3
DFTNTP	0	0	0	0	0	17	0	0	17	6

PEN	134	0	0	0	0	0	0	0	134	4

Source: Counted from simulation results

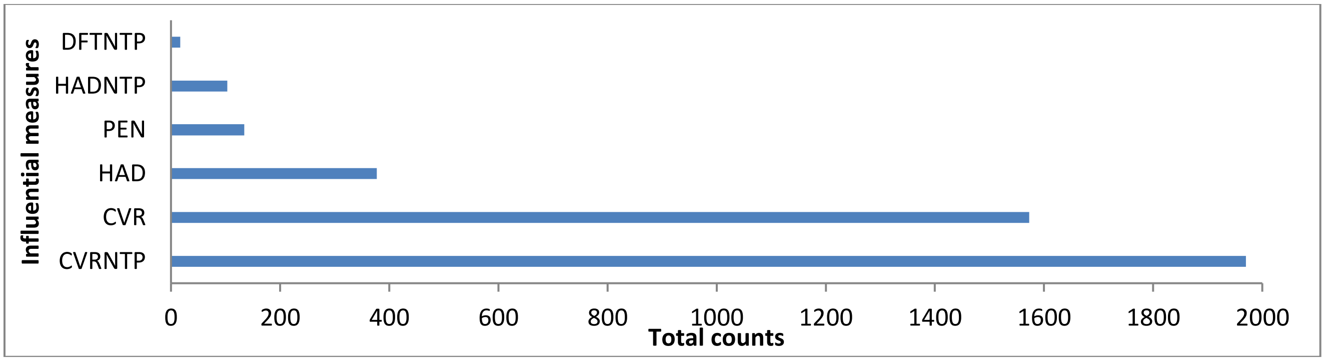

Figure 1: Graphical illustration of the best influential measures that correctly detect influential points when counted overall specifications.

4.1 Application; Longley data

Longley data used by Longley (1967) was employed in this study, the regression equation is defined as:

$$y = \theta_1 w_1 + \theta_2 w_2 + \theta_3 w_3 + \theta_4 w_4 + \theta_5 w_5 + \theta_6 w_6 \qquad (55)$$

where y is the total derived employment, x_1 is the gross national product implicit price deflator, w_2 is the gross national product, w_3 is unemployment, w_4 is the size of armed forces, $w_5$is the non-institutional population 14 years of age and over and w_6 is the time. Meanwhile, (Walker and Birch, 1988) affirmed that the scaled condition number of the data is 43.275. In the same vein, very many researchers have used this data to identify influential points, such as; ([59], [16], [29], [27], [64] , [57] , [33], and [38]. The value of parameter k used in this study is the one used by [64], [57] and [34], which was computed as 0.0012.

Table 9: Most five influential points detected by the existing influential measures compared with the proposed ones in NTP using Longley data sets.

		Influential Diagnostic based on OLS Estimator (Existing Method)		
Author(s)	**Year**	**Influential Measure**	**Cases identified in order**	**Method**
Cook	1977	CKD	5, 16, 4, 10, 15	Cook's distance based on OLS
Welsch and Kuh,	1977	DFT	5, 16, 10, 15, 4	DFFITs based on OLS
Hadi	1992	HAD	10, 5, 4, 16, 15	Hadi's measure based on OLS
Atkinson	1985	ATK	5, 16, 4, 10, 15	Atkinson measure based on OLS
Belsley *etal.*	1980	CVR	8, 2, 12, 9, 11	COVRATIO based on OLS
Pena	2005	PEN	16, 15, 14, 1, 12	Pena's Statistic based on OLS
		Influential Diagnostic based on NTP Estimator (Proposed Method)		
	Influential Measure		**Cases identified in order**	**Method**
	CKDNTP	Proposed	16, 10, 6, 4, 13	Cook's distance in NTP estimator
	DFTNTP	Proposed	10, 9, 11, 1, 8	DFFITs in the distance in NTP estimator
	ATKNTP	Proposed	16, 10, 6, 4, 2	Atkinson measure in NTP estimator
	CVRNTP	Proposed	12, 2, 9, 3, 8	COVRATIO in NTP estimator
	HAD_NTP	Proposed	16, 10, 2, 6, 4	Hadi's measure in the NTP estimator
	PEN_NTP	Proposed	16, 15, 14, 5, 7	Pena's Statistic in NTP estimator

5. Summary

The identification of influential points in linear regression is essential to avoid distorted inferential conclusions about regression coefficients. Existing diagnostic measures, primarily based on OLS, are limited by their dependence on the basic assumptions of linear regression, such as the absence of multicollinearity and outliers. This study introduced new diagnostic measures—DFFITs, Cook's D, COVRATIO, Hadi's measure, Pena's Statistic, and Atkinson Statistic using the New Two-Parameter estimator that can address multicollinearity problems. Simulation studies with 1,000 replications under various conditions, including levels of outliers, percentage of outliers, multicollinearity levels, error variances, and sample sizes, as well as applications to real-life data, revealed the superior performance of the proposed measures. The influential measures identified different percentages of influential point at different categories of the percentage of detection. Among them, CVRNTP consistently achieved 100% detection rates, outperforming existing OLS-based measures and demonstrating the highest detection counts of 1970. The proposed measures identified additional influential points that OLS-based measures missed, particularly under conditions of multicollinearity and outliers such as CKDNTP identified cases 6 and 13 more, DFTNTP identified cases 9, 11, 1 and 8, CVRNTP identified only case 3 more, cases 6 and 2 in this order were identified by Atkinson measure in NTP (ATKNTP), Hadi’s measures in NTP (HADNTP) identified cases 2 and 6 more meanwhile, Pena’s statistic in NTP (PENNTP) identified cases 5 and 7. The results confirm that the new diagnostic tools are robust and reliable, offering improved detection capabilities compared to existing methods.

6. Conclusions

This study highlights the critical role of identifying influential points in linear regression to ensure accurate inferential conclusions, particularly regarding regression coefficients. Existing diagnostic measures based on Ordinary Least Squares (OLS) often fail when the basic assumptions of linear regression, such as the absence of multicollinearity and outliers, are violated. To address this limitation, new diagnostic measures—DFFITs, Cook's D, COVRATIO, Hadi's measure, Pena's Statistic (PEN), and Atkinson Statistic were developed using the New Two-Parameter estimator designed to handle multicollinearity. Simulation studies and real-life data applications demonstrated that the proposed measures outperform their OLS-based counterparts in detecting influential points, especially under challenging conditions such as high multicollinearity, varying error variances, and the presence of outliers. Notably, the COVRATIO measure with NTP (CVRNTP) achieved 100% detection rates and exhibited the highest detection counts across all scenarios. The newly proposed measures also identified more influential points than

existing ones, confirming their robustness in practical applications. This research underscores the importance of addressing multicollinearity and outliers in regression models using appropriate estimators like NTP. Practitioners and policymakers can use these findings to decide whether to eliminate contaminated observations or adopt advanced diagnostic measures capable of handling such complexities, thereby improving the reliability of regression analysis.

List of Abbreviations

OLS: Ordinary Least Squares
CKD: Cook's D
DFT: DFFITs
HAD: Hadi's measure
ATK: Atkinson statistic
PEN: Pena's Statistic
NTP: New Two-Parameter estimator
DK: Dawoud-Kibria
MNTPE: Modified New Two-type parameter Estimator
PCR: Principal Component Regression
MM: MM-estimator
M: M – estimator
S: S-estimator
LMS: Least Median Squares estimator
LTS: Least Trimmed Squares estimator
LWS: Least Winsorized Squares
RMM: Ridge regression based on MM-estimator
DRGP: Robust Generalized Potentials for MM – estimator
DRGP-L: Diagnostic Robust Generalized Potentials for L-estimator
DRGP-LTS: Diagnostic Robust Generalized Potentials for LTS-estimator
DRGP-M: Diagnostic Robust Generalized Potentials for M –estimator
DRGP-MM: Diagnostic Robust Generalized Potentials for MM-estimator
MCRM: Multiple Circular Regression Model
CVRNTP: Covratio in New Two-Parameter estimator
HADNTP: Hadi’s measure in the New Two-Parameter estimator
DFTNTP: DFFITs in New Two-Parameter estimator
CVR: COVRATIO in OLS estimator
CKD: Cook’ D in OLS estimator
CKDNTP: Cook’s in New Two-Parameter Estimator
ATKNTP: Atkinson measure in NTP estimator
PENNTP: Pena's Statistic in NTP estimator

Suggestions for future research

The study only captured when multicollinearity is mitigated hence, the work can be extended when both multicollinearity and outliers are addressed simultaneously.

Declarations

Availability of data: Data used for simulation was generated using R-statistical programming codes, meanwhile, data used for application as real-life data will be made available upon request.

Acknowledgments: All peer Reviewers are appreciated for their constructive criticisms that actually improved the quality of this study.

References

[1] 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*

[2] 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*.

[3] Ajiboye, A.S., Adejumo, T. J. and Ayinde, K. (2017). A Study on Sensitivity and Robustness of Matched-Pairs Inferential Test Statistics to Outliers. *FUTA Journal of Research in Sciences, 13(2), 350-363.*

[4] Ajiboye, A.S., Adewuyi, E. Ayinde, K. and Lukman, A. F. (2016). A comparative study of some Robust Ridge and Liu Estimators. *Science world Journal, 11 (4), 1597 – 6343.*

[5] Alguraibawi, M., Midi, H. and Imon, A. H. M. R. (2015). A new Robust Diagnostic plot for classifying Good and Bad High leverage points in a multiple linear Regression Model. *Journal of Mathematical problems in Engineering. Doi:10.1155/2015/279472.*

[6] Alkasadi, N. A., Ibrahim, S., Abuzaid, A. H. M., Yusoff, M. I., Hamid, H., Waozhe, L. and Abdrasak, A. (2019). Outlier Detection in Multiple circular Degression Model using DFFITs statistics. *Sains Malaysiana 46 (7), 1557 – 1563. http://dx.doi.org/10.17576/.ism-2019-4807 – 25.*

[7] Atkinson, A. C. (1985). Plots, transformation and Regression: An introduction to graphical methods of diagnostic Regression analysis. Claredon press, 1985.

[8] Asar, Y. and Erisoglu, M. (2016). Influence diagnostics in Two-parameter Ridge Regression. *Journal of Data Science 14, 33-52. Doi: 10.6339/JDS. 201601_14(1).0003.*

[9] Ayinde, K., Adejumo, T. J. and Solomon, G. S. (2016). A study on sensitivity and Robustness of one sample test statistics to outliers. *Global Journal of Science Frontier Research: Mathematics and decision sciences, 16(6), 99-112.*

[10] Ayinde, K., Lukman, A. F. and Arowolo, O. T. (2015). Robust regression diagnostics of influential observations in linear regression model. *Open Journal of Statistics. 5, 1- 11.*

[11] Ayinde, K., Lukman, A. F., Alabi, O. O. and Bello, H. A. (2020). A new approach of principal component Regression Estimator with Applications to collinear data. *International Journal of Engineering Research and Technology, 13, 7, pp. 1616 – 1622.http://dx.doi.org/10.37624/IJERT/13.7.2020.1616 – 1622.*

[12] Bagheri, A., and Midi, H. (2009). Robust Estimations as a Remedy for multicollinearity caused by multiple High leverage points. *Journal of Mathematics and Statistics, 5 (4), 311 – 321.*

[13] Bagheri, A., Midi, H. and Imon, A. H. M. R. (2010). The effect of collinearity – influential observations on collinear Data set. A Montecarlo Simulation study. *Journal of Applied Sciences, 10, (18), 2086 -2093.*

[14] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). Regression Diagnostics: identifying influence Data and sources of collinearity. Wiley and sons, New York. *http://dx.doi.org/10.1002/0471725153.*

[15] Chang, X., and Yang, (2012). Combining two-parameter and principal component regression estimators. *Stat. papers, 53, 549 – 562.*

[16] Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics. 19; 15 – 18*

[17] 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*, *91*(17), 3678-3692.

[18] 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.*

[19] Dorugade, A. V. (2014). New ridge parameters for ridge regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences, 15, 94 – 99. Doi: 10.1016/j.jaubas.2013.03.005.*

[20] Draper, N. R. and John, J. (1981). Influential observations and outliers in regression, *Technometrics, 23(1), 21 – 26.*

[21] El-Fallah, M. and El-Salam, A. (2013). Alternative Outliers detection Procedures in Linear Regression Analysis: A comparative study. *International Journal of Mathematics and Statistics Studies, 2(1), 25-33.*

[22] Emami, H. and Emami, M. (2016). New influence diagnostics in ridge regression, *Journal of Applied Statistics, 43(3), 476-489. Doi: 10.1080/02664763.1070804.*

[23] Hadi, A. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics and Data Analysis. 14. 1-27. https://doi.org/10.1016/0167-9473192/90078-T.*

[24] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression. Biased Estimation for nonorthogonal problems. *Technometrics, 12(1), 55 – 67.*

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

[26] Hussein, E. (2021). Leverage and Influential observations on the Liu type estimator in the linear regression model with the severe collinearity. *Heliyon. Doi.org/10.1016/j.heliyon.2021.e07792.*

[27] Jahufer, A. (2013). Detecting Global influential observations in Liu Regression Model. *Open journal of statistics,3,5-11. http//dx.doi.org/10.4236/ojs.2013.31002.*

[28] Jahufer, A. and Jianbao, C. (2008). Assessing global influential observations in modified ridge regression. *Statistics and probability letters. 79 (2008). 513 -518. http://dx.doi.org / 10.1080/ 00401766. 1970.10488634.*

[29] Jahufer, A. and Jianbao, C. (2009). Assessing global influential observations in modified ridge regression. *Statistics and Probability Letters, 4(2), 513-518.*

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

[31] Kashif, M., Amanullah, M. and Aslam, M. (2018). Pena's statistic for the Liu regression. Journal of Statistic for the Liu regression. *Journal of Statistical Computation and Simulation, 88 (13), 2473 – 2488.*

[32] Kausar, T. Akbar, A. and Qasim, M. (2023). Influential Diagnostics for the Cox proportional hazards regression model: Method, Simulation and Application. *Journal of Statistical Computation and Simulation, 93(10), 1580 – 1600. 10.1080/00949655.2022.2145608.*

[33] Kashif, M., Ullah, M. A. and Aslam, M. (2019). Influential diagnostic with Pena's statistic for the Modified ridge regression. *Communication in Statistics-Simulation and Computation. Doi:10.1080/03610918.2019.1634204.*

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

[35] 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.*

[36] Longley, J. W. (1967). An appraisal of least squares programs for electronic computer from the point of view of the use. *Journal of American Statistical Association, 62, 819 – 841.*

[37] Lukman, A. F. and Ayinde, K. (2016). Detecting observations in Two-Parameter Liu-Ridge Estimator. *Journal of Data Science. 207218, Doi:10.6339/JDS.201804_16(2).0001.*

[38] Lukman, A. F. and Ayinde, K. (2018). Detecting influential observations in Two-parameter Liu-Ridge Estimator. *Journal of Data science, 16(2),0001, 201 -218. Doi:10.6339/JDS.201804.*

[39] 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.*

[40] Lukman, A. F., Ayinde, K., Binuomote, S. and Onate, A. C. (2019). Modified ridge -type estimator to combat multicollinearity: Application to Chemical data. *Journal of Chemometrics. Doi:10.1002,cem.3125.*

[41] Lukman, A. F., Ayinde, K., Okunola, A. O., Akanbi, O. B. and Onate, C. A. (2018). Classification-Based Ridge Estimation Techniques of Alkhamisi Methods. *Journal of probability and Statistical Sciences. 16(2), 2018. 165 – 181.*

[42] Meloun, M. and Militky, J. (2001). Detection of Single influential points in OLS regression model building. *Analytical Chimica Acta. Doi: 1.1016/50003.267(01)01040-6. 169 -191.*

[43] Midi, H. and Zahari, M. (2007): A simulation study on Ridge Regression Estimators in the presence of outliers and multicollinearity. *Journal of Teknologi, 47 (l). 59 – 74.*

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

[45] Pearson, K. (1901). On lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine, Series 6, 2(11), 559 – 572.*

[46] Pena, D. (2005). New Statistic for influence in linear regression. *Technometrics,47(1), Doi: 10.1198/004017004000000662, 47: 1-12.*

[47] Rao, C. R. (1973). Linear Statistical Inference and its applications, 22: John Wiley and Sons.

[48] Rosner, B. (1983). Percentage points for a generalized ESD many – outlier detection. New York.

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

[50] Rousseeuw, P. J. and Driessen, V. K. (1998). Computing LTS regression for large data sets, Technical Report University of Antwerp submitted.

[51] Rousseeuw, P. J. and Leroy, A. M. (1987). Robust Regression and Outlier Detection, Wiley, New York, USA.

[55] Rousseeuw, P. J. and Yohai (1984). Robust regression by means of S-estimator. In W. H. J. Frank and D. Martin; Robust and nonlinear Time series Analysis, Springer-verlag, New York, 256 – 272.

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

[57] 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.*

[58] Uzuke, C. A. and Ezeilo, I. C. (2021). On identifying influential observations in the presence of multicollinearity. *Open Journal of Statistics, 290 – 302. DOI: 10.423610js.2021.112016.*

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

[60] Welsch, R. E. (1982). Influence function and regression diagnostics. Modern Data Analysis. New york. Academic Press.

[61] Welsch, R. E. and Kuh, E. (1977). Linear regression diagnostics. Technical Report 923-77. Sloan school of management, Massachusetts Institute of Technology

[62] 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.*

[63] Yang, H. and Chang, X. (2012). Combining two-parameter and principal component regression estimators. Stat. papers, 53, 549 – 562.

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

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

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

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

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.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)