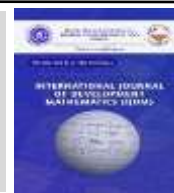




INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

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

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



Generalized Cook's Distance and DFFITS for the New Biased-Based Estimation in the Presence of Multicollinearity and Outliers

Olanrewaju O. Oladiran^{a*}, Timothy O. Olatayo^a, and Abass I. Taiwo^a.

^aDepartment of Statistics, Olabisi Onanbanjo University Ago-Iwoye, Nigeria

ARTICLE INFO

Article history:

Received 20 June 2026

Received in revised form 30 July 2026

Accepted 10 August 2026

Keywords:

Influence Diagnostics, Multicollinearity, Outliers, New Biased-Based (NBB), Cook's Distance and DFFITS

MSC 2020 Subject classification:

62P05

ABSTRACT

Regression analysis is a powerful tool for modelling relationships between variables, but its reliability hinges on meeting key assumptions of the classical linear regression model. When multicollinearity and outliers are simultaneously present, traditional estimation techniques such as Ordinary Least Squares (OLS) become unreliable, and classical influence diagnostics like Cook's Distance and DFFITS may fail to detect influential observations accurately. To address this gap, this study develops generalized versions of Cook's Distance and DFFITS tailored to the New Biased-Based (NBB) estimator, a two-parameter estimator designed to mitigate multicollinearity. Using case-deletion and approximate analytical approaches, we derive influence diagnostics that incorporate the structure of the NBB estimator. The proposed methods are then applied to economic and manufacturing datasets known to exhibit both multicollinearity and outliers. Results demonstrate that the new diagnostic measures, namely; Cook's D in NBB (CD_NBB) and DFFITS in NBB (DF_NBB) are effective in identifying influential observations, showing strong alignment with and, in some cases, improvements over existing robust and biased diagnostic techniques. These findings suggest that the NBB-based influence diagnostics offer a valuable addition to the toolkit for regression analysis under complex data conditions.

1 Introduction

Regression analysis is a core statistical method used to model relationships between a dependent variable and one or more independent variables. It underpins estimation, prediction, and causal inference across disciplines such as economics, engineering, medicine, and the social sciences. However, the reliability of regression outcomes depends on meeting key assumptions of the classical linear regression model. When these assumptions are violated particularly due to multicollinearity (strong correlation among predictors) and outliers (extreme or influential data points) estimates may become biased and inference invalid (Belsley *et al.* (1980). Multicollinearity increases the variance of Ordinary Least Squares (OLS) estimates, leading to unstable coefficients. At the same time, outliers can distort results and raise prediction errors. OLS is particularly vulnerable to these issues (Rousseeuw & Leroy, 1987). To address them, researchers have proposed biased estimators such as Ridge (Hoerl & Kennard, 1970) and (Liu, 1993) to manage multicollinearity, and robust estimators like M-, MM-, and S-estimators to reduce the impact of outliers.

It has been noted that biased estimators lack robustness, and at the same time robust methods can be inefficient under collinearity. This dual challenge also complicates the detection of influential observations, as traditional diagnostics like Cook's Distance and DFFITS developed based on classical method such as OLS may be distorted or masked under these conditions (Chatterjee and Hadi, 1986). Over the years, hybrid approaches have emerged, Kibria (2003) introduced Ridge-type Liu estimators to handle moderate multicollinearity problem. Lukman and Ayinde (2018) combined M-estimation with Ridge regression to enhance robustness. Ayinde and Adebayo (2020) developed robust

*Corresponding author. Tel.: +2347033886255

E-mail address: olanrewajuoladiranresearchstat@gmail.com (Oladiran, O. O.)

<https://doi.org/10.62054/ijdm/0303.210>

versions of the Liu estimator to address both problems simultaneously. Robust diagnostics have also been developed to complement these estimators.

Classical diagnostic tools like Cook's Distance, DFBETAS, DFFITs (Cook, 1977; Welsch & Kuh, 1977) often fail under high multicollinearity or when outliers are present. High leverage, while useful, can be misleading without considering residuals. Montgomery, *et al.* (2012). Chatterjee and Hadi (1988) noted that multicollinearity can obscure the influence of observations, prompting the use of condition indices and variance-decomposition proportions (Belsley *et al.* 1980).

To overcome these challenges, Ridge-based diagnostics were developed to stabilize influence measures under collinearity (Lawless and Wang, 1976), while robust methods such as forward search (Atkinson and Riani, 2000) and robust covariance matrices (Rousseeuw and van Zomeren, 1990) improved detection under outliers. Several authors have explored diagnostics tailored to biased estimators. Jahufer (2013) and Asar and Erisoglu (2016) proposed diagnostics for two-parameter Ridge estimators. However, several estimators such as the NBB by Sakalliglu and Kaciranlar (2008) are yet to be considered for influential diagnostic tool. In the same vein, numerous approaches have been developed for identifying influential data points in regression models, particularly under the complicating conditions of multicollinearity and outliers. Rahman *et al.* (2012) proposed a novel methodology for detecting outliers in multiple linear regression by determining whether atypical observations originate from distinct populations using mathematical criteria.

Traditional influential diagnostics such as Cook's distance (Cook, 1977), DFFITs (Welsch and Kuh, 1977), and Welsch's distance (Welsch, 1982) remain foundational tools for assessing the impact of extreme values or high-leverage points on regression coefficients. DFBETAS, for instance, quantifies the influence of a single observation on each regression coefficient. Meloun and Milityk (2001) conducted an in-depth study on the detection of single influential points in OLS regression, offering a synthesis of diagnostic measures and exploratory tools. Leiva *et al.* (2006) extended diagnostic methods to log-Birnbaum-Saunders regression models with censored data, introducing two deviance-based residuals tailored for outlier detection.

Jahufer and Jianbao (2008) proposed the Modified Ridge Regression (MRR) framework as an enhancement for identifying influential observations in the presence of multicollinearity. High-leverage observations often exacerbate multicollinearity, prompting the need for robust estimation procedures. Bagheri and Midi (2009) introduced the Diagnostic Robust Generalized Potentials (DRGP) for MM-estimators, demonstrating superior performance in identifying outliers in multicollinear settings. Subsequently, Bagheri *et al.* (2010) evaluated the impact of collinearity-influential observations using Monte Carlo simulations. Jahufer (2013) contributed a deletion-based method to detect influential cases in Liu-type regression estimators. Ayinde *et al.* (2015) comparatively evaluated classical and robust diagnostic methods including Cook's D, Welsch-Kuh distance, DFFITs, DFBETAS, MM-estimators, Least Trimmed Squares (LTS), and S-estimators based on their root mean square error (RMSE) performance on empirical data.

Asar and Erisoglu (2016) proposed a generalized diagnostic framework based on a two-parameter ridge-type estimator, extending classical influence measures and deriving corresponding deletion formulas. Emami and Emami (2016) utilized Pena's statistic to assess influential observations in ridge regression settings. Lukman and Ayinde (2016) presented an approximate deletion diagnostic for detecting influential points within the two-parameter Liu-Ridge regression framework, designed to simultaneously address multicollinearity and outliers.

Peng *et al.* (2016) proposed the Generalized Standardized Corrected Pearson Residuals (GSCPR) for detecting multiple leverage points in generalized linear models. Lukman *et al.* (2018) further modified Pena's statistic for biased estimation settings, while Kashif *et al.* (2018) introduced a new influence measure tailored to the Liu estimator. In the context of multiple circular regression models, Alkasadi *et al.* (2019) employed DFFITs for effective outlier detection. Hussein (2021) extended classical diagnostics by developing Liu-type versions of Cook's D and DFFITs to handle high multicollinearity in real-world datasets. Uzuke and Ezeilo (2021) compared the efficacy of diagnostic tools including Cook's D, DFFITs, Covratio, Hadi's measure, and DFBETAS under varying ridge parameters, concluding that DFFITs, DFBETAS, and Covratio are more effective when multicollinearity is addressed.

Adejumo *et al.* (2025) developed diagnostic measures tailored to the KL estimator in situations where only multicollinearity is present. In a parallel study, the same authors evaluated diagnostic tools within the framework of the newly proposed NTP estimator under joint multicollinearity and outlier conditions. Simulation results and empirical applications revealed that Covratio performed most reliably within the NTP framework. This study aims to

further develop influence diagnostics tailored to estimator that mitigate multicollinearity but remain sensitive to outliers. Specifically, it proposes approximate deletion formulas for Cook's Distance and DFFITs designed for the estimator. The goal is to evaluate their effectiveness in identifying influential observations using real-life data set, comparing results with those of existing methods.

2.1 Influential Diagnostics in Ordinary Regression model

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 (CD) Diagnostic Measure

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

$$CD_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (\hat{\beta}_{(i)} - \hat{\beta})}{p \hat{\sigma}^2} \quad (1)$$

Where $\hat{\beta}_{(i)}$ and $\hat{\beta}$ 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 (2.16) can be defined as:

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

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

$$t_i = \frac{\ell_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}}. \quad (3)$$

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.

Cook (1977) suggested that observations for which $CD_i > 1$ warrant attention.

2.1.2 DFFITs (DF) Influential Diagnostic

Welsch and Kuh (1977) proposed DFFITs diagnostic influential measure which is defined as the deletion influence of i th 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 i th data is deleted divided by the estimated standard deviation of the fit at that point. The statistic can be expressed as:

$$DFS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}_{(i)}^2 h_{ii}}}, \quad i = 1, 2, 3, \dots, n \quad (4)$$

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 Lukman and Ayinde (2018). Equation (4) can also be expressed as:

$$DFS_i = \frac{x_i [\hat{\beta} - \hat{\beta}_{(i)}]}{S(x_i \hat{\beta})}. \quad (5)$$

where $S(x_i\hat{\beta})$ is an estimator of standard error of the fitted values, x_i is the i th row of the X matrix, $\hat{\beta}_{(i)}$ is the least squares estimator of θ when the i th case is deleted in fitting the linear regression model. Further algebraic expression, equation (5) can be expressed as:

$$DFS_i = \left(\frac{h_{ii}}{1-h_{ii}} \right)^{\frac{1}{2}} \frac{\ell_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}, \quad (6)$$

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

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.

Jahufer (2013) suggested that observations for which $|DFS_i| > 2\sqrt{\frac{p}{n}}$ warrant attention for large data sets. Also, if the absolute value of DFS_i exceeds 1 for small to medium data sets, such observation is influential

3 Proposed Influential Diagnostics in New Biased-Based (NBB) Estimator

As a means of mitigating the effect of multicollinearity in regression analysis, Sakalliglu and Kaciranlar (2008) developed a new estimator named the New Biased-Based (NBB) estimator. The estimator is defined as:

$$\hat{\beta}_{(k-d)}^{NBB} = (X^T X + I)^{-1} (X^T X + (k+d)I) \hat{\beta}_{ORR}.$$

where $\hat{\beta}_{ORR} = (X^T X + kI)^{-1} X^T Y$ such that, k and d are the estimated biasing parameters for the NBB-estimator which are expressed as:

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

such that $0 < k < 1$, λ_i is the i th Eigen value of $X^T X$, $\hat{\alpha}_i = Q^T \hat{\beta}_{OLS}$ and $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y})^2}{n-p}$ which is the MSE of

OLS with complete data.

Several versions of Cook's D and DFFITs measures in one-parameter estimators have been developed, but limited studies have considered Two-Parameter estimators. One of the Two-Parameter estimators that are yet to be considered is the NBB estimator. On this note, the generalized version of Cook's D and DFFITs including their approximate case deletion formulas are hereby developed.

Hence, the idea employed in this study is that the estimator and the hat matrix of the fitted value of the above estimator were used in replacement of the OLS estimator and the hat matrix of the fitted value used in the conventional influential measures mentioned above.

The fitted value can be estimated as follows:

$$\hat{Y}_i^{NBB} = X \hat{\beta}_{(k-d)}^{NBB},$$

$$\begin{aligned}\hat{Y}_i^{NBB} &= X(X^T X + I)^{-1}(X^T X + (k+d)I)\hat{\beta}_{ORR}, \\ \hat{Y}_i^{NBB} &= X(X^T X + I)^{-1}(X^T X + (k+d)I)(X^T X + kI)^{-1}X^T Y, \\ \hat{Y}_i^{NBB} &= H^{NBB}Y.\end{aligned}$$

where $H^{NBB} = X(X^T X + I)^{-1}(X^T X + (k+d)I)(X^T X + kI)^{-1}X^T$, is the hat matrix of the fitted value for the NBB estimator and it is equivalent to the hat matrix of the fitted value in OLS estimator when $k = 0$ and $d = 1$. In this case, H^{NBB} is not idempotent; therefore, it is not a projection matrix. The i th diagonal element of the H^{NBB} can be expressed as:

$$h_{ii}^{NBB} = x_i(X^T X + I)^{-1}(X^T X + (k+d)I)(X^T X + kI)^{-1}x_i^T.$$

Such that x_i is the i th row of the matrix X and the i th fitted value written in terms of element H^{NBB} is defined as:

$$\hat{Y}_i^{NBB} = \sum_{i=1}^n h_{ii}^{NBB} Y_i. \quad (7)$$

Differentiating $\hat{Y}_i^{NBB}$ partially with respect to Y_i ;

$$\frac{\partial \hat{Y}_i^{NBB}}{\partial Y_i} = h_{ii}^{NBB}. \quad (8)$$

Such that, the i th New Biased-Based error term is given as:

$$e_i^{NBB} = Y_i - \hat{Y}_i^{NBB}, \quad (9)$$

$$e_i^{NBB} = (1 - h_{ii}^{NBB})Y_i. \quad (10)$$

3.1 Cook's D measure in New Biased-Based (NBB) Estimator

In respect to the procedures of conventional Cook's D proposed based on Ordinary Least Squares by Cook (1977), the generalized version of Cook's D in NBB estimator is hereby proposed and defined as:

$$CD_i^{NBB} = \frac{(\hat{\beta}_{(k-d)}^{NBB} - \hat{\beta}_{(i)(k-d)}^{NBB})^T (X^T X) (\hat{\beta}_{(k-d)}^{NBB} - \hat{\beta}_{(i)(k-d)}^{NBB})}{pS^2}, \quad (11)$$

The equivalent version of D_i^{NBB} when $k=0$ and $d=1$ is given as:

$$CD_i^{*NBB} = \frac{(\hat{\beta}_{(k-d)}^{NBB} - \hat{\beta}_{(i)(k-d)}^{NBB})^T E(X^T X) E(\hat{\beta}_{(k-d)}^{NBB} - \hat{\beta}_{(i)(k-d)}^{NBB})}{pS^2}. \quad (12)$$

where $E = (X^T X + I)^{-1}(X^T X + (k+d)I)(X^T X + kI)^{-1}$.

Therefore, based on $\text{var}(\hat{\beta}_{(k-d)}^{NBB}) = \sigma^2 E(X^T X)^{-1} E$ and as a result of the dependency of the NBB estimator, the design matrix X with the i th row deleted needs to be rescaled before $\hat{\beta}_{(k-d)}^{NBB}$ is computed. Hence, the two measures CD_i^{NBB} and CD_i^{*NBB} can not be written as a function of leverage and residual.

Also, following the procedures of conventional Cook's D measure and by applying Sherman Morrison–Woodbury theorem by Rao (1973), the Cook's D in NBB estimator can further be expressed as:

$$CD_i^{NBB} = \frac{(e_i^{NBB})^2}{pS^2} \left(\frac{h_{ii}^{NBB}}{1 - h_{ii}^{NBB}} \right).$$

where h_{ii}^{NBB} is the diagonal element of the hat matrix H^{NBB} and S^2 is the MSE when the data is complete. The cut-off point is $CD_i^{NBB} > 1$, this implies that the observation that its $D_i^{NBB} > 1$ when sample size (n) is large is said to be influential and thereby warrants attention.

3.2 DFFITs in New Biased-Based (NBB) Estimator

Following the conventional DFFITs measure proposed based on Ordinary Least Squares estimator, the generalized version of DFFITs in NBB estimator is hereby developed and denoted as:

$$DFS_i^{NBB} = \frac{x_i (\hat{\beta}_{(k-d)}^{NBB} - \hat{\beta}_{(i)(k-d)}^{NBB})}{SE(x_i \hat{\beta}_{(k-d)}^{NBB})}. \quad (13)$$

where $SE(x_i \hat{\beta}_{(k-d)}^{NBB}) = s \left(x_i E(X^T X)^{-1} E x_i^T \right)^{\frac{1}{2}}$,

such that $E = (X^T X + I)^{-1} (X^T X + (k+d)I) (X^T X + kI)^{-1}$

This is the standard error of the NBB estimator fitted values. It can also be expressed as:

$$SE(x_i \hat{\beta}_{(k-d)}^{NBB}) = s_{(i)} \sqrt{\sum_{j=1}^n (h_{ij}^{NBB})^2}. \quad (14)$$

where $s_{(i)} = \sqrt{\frac{(n-p)s^2 - e_i^2}{n-p-1(1-h_{ii})}}$ and h_{ij}^{NBB} refers to the ij th element of the H^{NBB} and s is the unbiased OLS estimator of σ .

Also, DFS_i^{NBB} can further be expressed as:

$$DFS_i^{NBB} = \frac{|\hat{Y}_i^{NBB} - \hat{Y}_{(-i)}^{NBB}|}{S_{(-i)NBB} \sqrt{h_{ii}^{NBB}}}. \quad (15)$$

where $\hat{Y}_i^{NBB}$ and $\hat{Y}_{(-i)}^{NBB}$ are the NBB-estimator predicted values of response variable Y when data is complete and when the i th observation is deleted, $S_{(-i)NBB}$ is the NBB estimator standard error estimated without the i th observation in question and h_{ii}^{NBB} is the diagonal elements of the hat matrix H^{NBB} .

The observation for which $DFS_i^{NBB} / > 2\sqrt{\frac{p}{n}}$ for large data, such observation is considered to be influential.

3.3 Approximate Case Deletion formulas for NBB Estimator

If the i th row is deleted from $\hat{\beta}_{(k,d)}^{NBB}$, therefore $\hat{\beta}_{(i)(k-d)}^{NBB}$ which is the estimator of the New Biased-Based when the i th observation has been removed and it can be written as:

$$\hat{\beta}_{(i)(k-d)}^{NBB} = (X_{(i)}^T X_{(i)} + I)^{-1} (X_{(i)} X_{(i)} + (k+d)I) (X_{(i)}^T X_{(i)} + kI)^{-1} X_{(i)}^T Y_{(i)}. \quad (16)$$

Such that $X_{(i)}$ is the Matrix X without the i th row. $X_{(i)}$ is scaled so that $X_{(i)}^T X_{(i)}$ is in correlation form.

$\hat{\beta}_{(i)(k-d)}^{NBB}$ is approximately obtained by applying Sherman-Morrison-Woodbury (SMW) theorem. It is therefore obtained as follows:

$$\begin{aligned} \hat{\beta}_{(i)(k-d)}^{NBB} &= (X^T X - x_i^T x_i + I)^{-1} (X^T X - x_i^T x_i + (k+d)I) \hat{\beta}_{ORR} \\ \hat{\beta}_{(i)(k-d)}^{NBB} &= (S_k - x_i^T x_i)^{-1} (X^T X - x_i^T x_i + (k+d)I) \hat{\beta}_{ORR}, \end{aligned} \quad (17)$$

where, $S_k = (X^T X + I)^{-1}$,

$$(S_k - x_i^T x_i)^{-1} = S_k^{-1} + \frac{S_k^{-1} x_i^T x_i S_k^{-1}}{1 - x_i S_k^{-1} x_i^T},$$

and

$$\begin{aligned} \hat{\beta}_{(i)(k-d)}^{NBB} &= \left(S_k^{-1} + \frac{S_k^{-1} x_i^T x_i S_k^{-1}}{1 - x_i S_k^{-1} x_i^T} \right) (X^T X - x_i^T x_i + (k+d)I) \hat{\beta}_{ORR}, \\ \hat{\beta}_{(i)(k-d)}^{NBB} &= \left(S_k^{-1} + \frac{S_k^{-1} x_i^T x_i S_k^{-1}}{1 - V_{ii}} \right) (X^T X - x_i^T x_i + (k+d)I) \hat{\beta}_{ORR}, \end{aligned} \quad (18)$$

where, $V_{ii} = x_i S_k^{-1} x_i^T$,

By expanding (18),

$$\hat{\beta}_{(i)(k-d)}^{NBB} = (X^T X + (k+d)I)S_k^{-1} + \hat{\beta}_{ORR} \left[\frac{(X^T X)x_i^T x_i S_k^{-1} + (k+d)IS_k^{-1} + S_k^{-1}x_i^T x_i S_k^{-1}}{1-V_{ii}} - \frac{x_i^T x_i S_k^{-1}x_i^T x_i S_k^{-1}}{1-V_{ii}} - x_i^T x_i S_k^{-1} \right] \hat{\beta}_{ORR}, \quad (19)$$

$$\hat{\beta}_{(i)(k-d)}^{NBB} = S_k^{-1}(X^T X + (k+d)I)\hat{\beta}_{ORR} + \frac{S_k^{-1}x_i^T}{1-V_{ii}} [(X^T X)x_i S_k^{-1} + (k+d)Ix_i S_k^{-1} - x_i^T x_i S_k^{-1} - x_i(1-V_{ii})]\hat{\beta}_{ORR} \quad (20)$$

where, $S_k^{-1}(X^T X + (k+d)I)\hat{\beta}_{ORR} = ((X^T X + I)^{-1}(X^T X + (k+d)I))\hat{\beta}_{ORR} = \hat{\beta}_{(k-d)}$,
Therefore,

$$\hat{\beta}_{(i)(k-d)}^{NBB} = \hat{\beta}_{(k-d)} + \frac{S_k^{-1}x_i^T}{1-V_{ii}} [x_i S_k^{-1}(X^T X + (k+d)I) - x_i^T x_i S_k^{-1}x_i(1-V_{ii})]\hat{\beta}_{ORR}, \quad (21)$$

$$\hat{\beta}_{(i)(k-d)}^{NBB} = \hat{\beta}_{(k-d)} - \frac{S_k^{-1}x_i^T}{1-V_{ii}} [-x_i S_k^{-1}(X^T X + (k+d)I) + x_i V_{ii} + x_i(1-V_{ii})]\hat{\beta}_{ORR}, \quad (22)$$

$$\hat{\beta}_{(i)(k-d)}^{NBB} = \hat{\beta}_{(k-d)} - \frac{S_k^{-1}x_i^T}{1-V_{ii}} [x_i - x_i S_k^{-1}(X^T X + (k+d)I)]\hat{\beta}_{ORR}. \quad (23)$$

Hence the approximate difference between $\hat{\beta}_{(i)(k-d)}^{NBB}$ and $\hat{\beta}_{(k-d)}$ is given as:

$$\hat{\beta}_{(k-d)}^{NBB} - \hat{\beta}_{(i)(k-d)}^{NBB} \cong \frac{S_k^{-1}x_i}{1-V_{ii}} [x_i(1 - S_k^{-1}(X^T X + (k+d)I))](X^T X + kI)^{-1} X^T Y. \quad (24)$$

where $S_k^{-1} = (X^T X + I)^{-1}$ and $V_{ii} = x_i^T (X^T X + I)^{-1} x_i$.

3.3.1 The approximate version for Cook's D in NBB-Estimator

The approximate version for Cook's distance in NBB estimator is hereby presented as:

$$D_i^{aNBB} = \frac{1}{pS^2} [J\hat{\beta}_{ORR}]^T (X^T X) [J\hat{\beta}_{ORR}]. \quad (25)$$

where

$$J = \frac{S_k^{-1}x_i}{1-V_{ii}} [x_i(1 - S_k^{-1}(X^T X + (k+d)I))].$$

3.3.2 The approximate version for DFFITs in NBB Estimator

The approximate version for DFFITs in NBB estimator can be expressed as:

$$DFS_i^{aNBB} = \left[\frac{V_{ii}}{1 - V_{ii}} \right] \frac{e_i^{NBB}}{S_{(i)} \sqrt{\sum_{j=1}^n (h_{ij}^{NBB})^2}}, \quad (26)$$

$$\text{where } S_{(i)} = \sqrt{\frac{(n-p)S^2 - \frac{e_i^2}{1-h_{ii}}}{n-p-1}} \text{ such that } V_{ii} = x_i^T (X^T X + I)^{-1} x_i, \text{ and } e_i^{NBB} = Y_i - \hat{Y}_i^{NBB}.$$

4 Application to Real-life

Real-life data were used to illustrate the performance of the influential diagnostics tools.

4.1 Application to Economic data

Several researchers, including Jahufer and Jianbao (2009), Jahufer (2013), Yasin and Murat (2016), Ullah *et al.* (2013), Kashif *et al.* (2019), Lukman and Ayinde (2018), and Longley (1967), have employed the economic dataset utilized in this study. The linear regression model for this dataset is given as:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + e_i \quad (27)$$

where:

y : total derived employment,

x_1 : gross national product (GNP) implicit price deflator,

x_2 : gross national product (GNP),

x_3 : unemployment rate,

x_4 : size of the armed forces,

x_5 : non-institutional civilian population aged 14 years and over,

x_6 : time (trend variable),

e_i : random error term.

Walker and Birch (1988) reported a scaled condition number (CN) of 43.275 for the dataset, indicating the presence of multicollinearity. Liu (1993) further substantiated this claim, noting variance inflation factor (VIF) values of 128.29, 103.43, and 70.87 for key predictors—well above the conventional threshold. This diagnosis was supported by Midi and Zahari (2007), who additionally identified the presence of influential outliers in the dataset. These findings underscore the challenges posed by both multicollinearity and data contamination in modeling this economic data.

Table 1: The five most influential cases detected by some proposed influential measures compared with some already existing influential measures using Economic data set.

Existing Methods			
Author(s)	Year	Influential cases in order	Method
Cook	1977	5, 16, 4, 10, 15	Cooks D based on OLS
Walker and Birch	1988	16, 10, 4, 15, 1	Cook's D in ridge regression
Shi and Wang	1999	10, 4, 15, 16, 1	Local influence in ridge regression
Jahufer	2013	15, 4, 1, 6, 16	Cooks D and DFFITs in Liu regression
		16, 10, 4, 6, 1	
		16, 10, 4, 6, 5, and	Influential points in Liu regression for
Ullah <i>et al.</i>	2013	16, 5, 4, 10, 15	d = 0.1, 0.5 and 0.9

Yasin and Murat	2016	16, 10, 6, 1 and 4	Influential points in Two-parameter Ridge
Lukman and Ayinde	2018	16, 10, 4, 5 and 15 3, 10, 11, 4 and 5 10, 3, 4, 15 and 16, 10, 4, 15, 5 and 16	DFFITs in Two-parameter Ridge_Liu
Kashif <i>et al.</i>	2018	15, 10, 5, 4 and 16	Pena's Statistic for the Liu regression for d = 0.1, 0.5, 0.9 and 1. 0
Kashif <i>et al.</i>	2019	16, 15, 5, 4 and 10	Modified Pena's Statistic for the Modified Ridge Regression for K = 0.0002
Dawoud and Hussein	2025	5, 15, 6, 13 and 1	DFFITs in Dawoud-Kibria
Proposed Methods			
CD_NBB	Proposed	4, 2, 9, 15, 16	Cook's D in New Biased Based with k=0.0012 and d= 0.5
DF-NBB	Proposed	16, 6, 13, 5, 2	DFFITs in New Biased Based with k=0.0012 and d= 0.5

Applying the proposed diagnostic methods to the economic dataset described in Section 4.1, the results presented in Table 1 demonstrate the effectiveness of the newly developed influential measures. Specifically, the Cook's D based on the New Biased-Based (NBB) estimator, denoted as CD_NBB, successfully identified cases 4, 2, 9, 15, and 16 as influential. These findings are consistent with several existing diagnostic methods, albeit in a different ranking order. The analysis employed biasing parameters $k=0.0012$ and $d=0.5$. Notably, cases 4, 15, and 16 were consistently identified across both the new and existing diagnostics, albeit with variations in their rankings aligning with findings reported by other researchers, as reflected in Table 1. Interestingly, CD_NBB also flagged cases 2 and 9 as influential cases not detected by the traditional methods in terms of ranking. Similarly, the new diagnostic measure DFFITs based on NBB, denoted as DF_NBB, identified cases 16, 6, 13, 5, and 2 as influential.

These results show partial overlap with existing methods, with the exception of cases 6 and 2, which were uniquely flagged by DF_NBB. Among these, case 16 emerged as the most consistently influential across nearly all methods examined, indicating its significant impact on the regression model. In conclusion, the newly developed diagnostic methods, particularly CD_NBB and DF_NBB, demonstrated comparable performance to existing techniques. They effectively identified key influential cases specifically observations 16, 4, 15, 13, and 2 reinforcing their robustness in detecting influential points, even under the presence of multicollinearity and potential outliers.

4.2 Application to Manufacturing data

The manufacturing dataset originally adopted by Ayinde *et al.* (2015) and subsequently utilized by various researchers is also employed in this study. The corresponding linear regression model for this dataset is specified as:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e_i \quad (28)$$

where:

y: value of manufactured output,

x_1 : value of imported intermediate goods,

x_1 : value of imported capital commodities,

x_1 : value of imported raw materials,

e_i : random error term.

The presence of severe multicollinearity within the dataset has been documented by Lukman *et al.* (2014), who reported variance inflation factor (VIF) values of 128.29, 103.43, and 70.87 values that were initially estimated by Hussein and Abdalla (2012). These results confirm that the regressors are highly correlated. Additionally, Hussein and Abdalla (2012) identified significant outliers in the data, indicating potential violations of the classical linear regression assumptions. In addressing these issues, this study adopts the shrinkage parameter values ($k=0.0012$,

$d=0.5$) previously applied by Yasin and Murat (2016) and Kashif *et al.* (2019), as they were found to improve estimator performance in the presence of multicollinearity.

Table 2: Most five influential cases detected by some proposed influential measures compared with some already existing influential measures using Manufacturing data.

Existing Methods			
Author(s)	Year	Influential points in order	Method
Cook	1977	15, 16, 31, 30 and 20	Cooks D in OLS
Ayinde <i>et al.</i>	2015	30, 31, 16, 21 and 14	Cooks D in M-estimator
Ayinde <i>et al.</i>	2015	15, 30, 31, 14 and 12	Cooks D in MM-estimator
Ayinde <i>et al.</i>	2015	15, 16, 31, 30 and 20	DFFITs based on OLS
Ayinde <i>et al.</i>	2015	31, 15, 30, 14 and 12	DFFITs in MM-estimator
Dawoud and Hussein	2025	15, 20, 29, 27 and 14	DFFITs in Dawoud-Kibria
Proposed Methods			
CD_NBB	Proposed	15, 16, 31, 30 and 20	Cook's D in New Biased Based Estimator with $k=0.0012$ and $d=0.5$
DF_NBB	Proposed	15, 20, 29, 27 and 14	DFFITs in New Biased Based Estimator with $k=0.0012$ and $d=0.5$

The proposed influential diagnostic measures were applied to the manufacturing dataset, which has been previously analyzed by several researchers including Lukman *et al.* (2014), Yasin and Murat (2016), and Kashif *et al.* (2019). As shown in Table 2, different existing influential measures identified varying sets of influential cases. For instance, DFFITs based on Ordinary Least Squares (OLS), as proposed by Ayinde *et al.* (2015), flagged cases 15, 16, 31, 30, and 20 as influential. Cook's D based on the M-estimator, also developed by Ayinde *et al.* (2015), identified cases 30, 31, 16, 21, and 14, while their DFFITs based on the MM-estimator identified cases 31, 15, 30, 14, and 12. In comparison, the newly developed influential measure Cook's D based on the New Biased-Based estimator (CD_NBB) successfully detected cases 4, 2, 9, 15, and 16 as influential. Similarly, the DFFITs based on the DK estimator introduced by Dawoud and Hussein (2025) identified cases 15, 20, 29, 27, and 14. Notably, the application of the newly proposed measures, CD_NBB and DF_NBB, to the manufacturing data revealed a strong alignment with the results from Ayinde *et al.* (2015) and Dawoud and Hussein (2025).

In fact, the influential cases identified by the new methods were largely consistent with those identified by the existing techniques and in some cases, matched the ranking order precisely. Overall, a comparative analysis of the diagnostic results suggests that cases 15 and 20 are consistently recognized as among the most influential observations across methods. This confirms that the newly developed diagnostics CD_NBB and DF_NBB not only perform competitively but also exhibit robustness comparable to established methods.

5 Conclusion

This study introduces novel influence diagnostic measures generalized and approximate versions of Cook's Distance and DFFITs within the framework of the New Biased-Based (NBB) estimator. These diagnostics are specifically designed to operate effectively under challenging conditions of multicollinearity and outliers, where classical methods often fall short. Theoretical derivations leverage the hat matrix and fitted values of the NBB estimator to redefine the influence measures, while approximate formulas based on the Sherman-Morrison-Woodbury theorem allow practical application without re-estimating the model for each case deletion.

Applications to real-world economic and manufacturing datasets confirm the utility of the proposed diagnostics. The NBB-based measures consistently identified influential observations that aligned with, and in some cases extended

beyond, those detected by conventional and robust methods. Notably, CD_NBB and DF_NBB demonstrated robustness and sensitivity across datasets, affirming their reliability and potential for broader application. In summary, the development of influence diagnostics for the NBB estimator fills an important methodological gap in regression analysis.

These tools provide researchers and analysts with improved capability to detect influential cases in the presence of both multicollinearity and outliers, thereby enhancing the robustness and interpretability of regression results. Hence, these measures will help statistics users decide whether to remove, retain, or otherwise address influential observations using robust methods when they are identified in the data set.

Acknowledgements: All peer reviewers are appreciated for their constructive criticisms toward the betterment of the study.

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

References

- Adejumo, T. J., Amin, M., Ayinde, K. and Adewuyi, E. T. (2025). Influential diagnostics in the Linear _model with Kibria-Lukman Estimator. *Communications in Statistics-Simulation and Computation*. <https://doi.org/10.1080/03610918.2025.2484628>.
- Alkasadi, A., Asar, Y., and Yildiz, O. (2019). Influence diagnostics in multiple circular regression models. *Communications in Statistics - Simulation and Computation*, 48(9), 2786–2799. <https://doi.org/10.1080/03610918.2018.1441171>
- Asar, Y., and Erisoglu, M. (2016). A new two-parameter estimator to combat multicollinearity. *Communications in Statistics - Theory and Methods*, 45(2), 555–565. <https://doi.org/10.1080/03610926.2013.841089>
- Atkinson, A. C., and Riani, M. (2000). *Robust diagnostic regression analysis*. Springer.
- Ayinde, K., and Adebayo, A. A. (2020). A new family of robust Liu-type estimators for regression analysis with outliers and multicollinearity. *Journal of Applied Statistics*, 47(1), 1–17. <https://doi.org/10.1080/02664763.2019.1616827>
- Ayinde, K., Lukman, A. F., and Abdulrahman, A. (2015). Diagnostic measures of influence in classical and robust regression estimators. *Journal of Modern Applied Statistical Methods*, 14(1), 240–258. <https://doi.org/10.22237/jmasm/1430453820>
- Bagheri, A., and Midi, H. (2009). Diagnostic robust generalized potentials for the detection of influential observations in linear regression. *Journal of Applied Statistics*, 36(5), 509–520. <https://doi.org/10.1080/02664760802320479>
- Bagheri, A., Midi, H., and Karimi, R. (2010). Collinearity-influential observations in regression: A Monte Carlo

- study. *Statistical Papers*, 51(4), 859–871. <https://doi.org/10.1007/s00362-008-0158-6>
- Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley.
- Chatterjee, S., and Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3), 379–393. <https://doi.org/10.1214/ss/1177013622>
- Chatterjee, S., and Hadi, A. S. (1988). *Sensitivity analysis in linear regression*. Wiley.
- Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1), 15–18. <https://doi.org/10.1080/00401706.1977.10489493>
- Dawoud, I. and Hussein, E. (2025). Detection of Influential Observations for the regression model in the presence of multicollinearity: theory and methods. *Communications in Statistics - Theory and Methods*. Doi [10.1080/03610926.2024.244910](https://doi.org/10.1080/03610926.2024.244910).
- Emami, A., and Emami, M. (2016). Influence diagnostics using Pena's statistic in ridge regression. *Journal of Statistical Computation and Simulation*, 86(12), 2343–2357. <https://doi.org/10.1080/00949655.2015.1121874>
- Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>
- Hussein, E. E. (2021). Modified diagnostic tools for Liu-type estimators under multicollinearity. *International Journal of Statistics and Applications*, 11(1), 1–9.
- 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.
- Jahufer, M. (2013). On detecting influential observations in regression models based on robust and biased estimators. *Communications in Statistics - Theory and Methods*, 42(14), 2544–2557. <https://doi.org/10.1080/03610926.2011.651453>
- Jahufer, M., and Jianbao, W. (2008). Modified ridge regression (MRR) for detecting influential observations in presence of multicollinearity. *Pakistan Journal of Statistics*, 24(2), 123–132.
- Kashif, M., Amanullah, M. and Aslam, M. (2018). Pena's statistic for the Liu regression. *Journal of Statistical Computation and Simulation* 88 (13): 2473 – 2488.
- 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](https://doi.org/10.1080/03610918.2019.1634204).

- Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics—Simulation and Computation*, 32(2), 419–435. <https://doi.org/10.1081/SAC-120017499>
- Lawless, J. F., and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics - Theory and Methods*, 5(4), 307–323. <https://doi.org/10.1080/03610927608827334>
- Leiva, V., Sanhueza, A., and Silva, M. (2006). Diagnostics in log-Birnbaum–Saunders regression models with censored data. *Computational Statistics & Data Analysis*, 51(3), 2214–2231. <https://doi.org/10.1016/j.csda.2006.06.003>
- Liu, K. (1993). A new class of biased estimate in linear regression. *Communications in Statistics - Theory and Methods*, 22(2), 393–402. <https://doi.org/10.1080/03610929308831024>
- 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.
- Lukman A. F., Arowolo O. and Ayinde K. (2014). Some robust ridge regression for handling multicollinearity and Outliers. *International Journal of Science and Applied Research (IJSBAR)*, 16, 192. <https://core.ac.uk/download/pdf/249333933.pdf>.
- Lukman, A. F., and Ayinde, K. (2016). Approximate deletion diagnostic for influential points in two-parameter Liu–Ridge regression. *Pakistan Journal of Statistics and Operation Research*, 12(4), 581–597. <https://doi.org/10.18187/pjsor.v12i4.1493>
- Lukman, A. F., and Ayinde, K. (2018). Performance of some ridge-type estimators when outliers are present. *Sri Lankan Journal of Applied Statistics*, 19(4), 299–316. <https://doi.org/10.4038/sljastats.v19i4.8005>
- 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.
- Meloun, M. and Militky, J. (2001). Detection of Single influential points in OLS regression model building. *Analytical Chimica Acta*. Doi:10.1016/S0003-2670(01)01040-6. 169 -191.
- Midi, H., and Zahari, M. (2007). Detecting multivariate outliers in regression. *Statistics in Transition*, 8(3), 435–448.
- Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.
- Peng, Y., Qian, Z., and Sun, Y. (2016). Generalized standardized corrected Pearson residuals in GLMs. *Journal of Statistical Computation and Simulation*, 86(7), 1396–1410. <https://doi.org/10.1080/00949655.2015.1051435>
- Rahman, M. A., Haque, M. M., and Akhter, S. (2012). Detection of outlying cases in multiple linear regression using

- population comparison. *Journal of Statistical Research*, 46(2), 221–231.
- Rao, C. R. (1973). *Linear Statistical Inference and its applications*, 22: John Wiley and Sons.
- Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley.
- Rousseeuw, P. J., and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), 633–639. <https://doi.org/10.2307/2290564>
- Sakallıoglu, S., and Kaciranlar, S. (2008). A new class of biased estimators for the linear regression model. *Hacettepe Journal of Mathematics and Statistics*, 37(1), 41–49.
- Shi, L. and Wang, X. (1999). Local Influence in Ridge Regression. *Compt Stat and Data Anal.* 31: 341-353.
- Ullah, M. I., Ahmed, N., and Ahmed, S. (2013). Influential observations in Liu regression models. *Pakistan Journal of Statistics*, 29(3), 343–356.
- Uzuke, R. O., and Ezeilo, M. A. (2021). Comparative performance of influence measures under ridge regression. *Nigerian Journal of Statistics*, 38(1), 85–104.
- Walker, A. M., and Birch, H. G. (1988). Ridge regression diagnostics and collinearity. *Journal of Econometrics*, 37(1), 53–70.
- Welsch, R. E. (1982). Influence functions for diagnostics. In *Modern Data Analysis* (pp. 149–167). Academic Press.
- Welsch, R. E., and Kuh, E. (1977). *Linear regression diagnostics*. NBER Technical Paper Series.
- Yasin, M., and Murat, A. (2016). Ridge-type estimators under collinearity and outlier conditions. *Hacettepe Journal of Mathematics and Statistics*, 45(3), 781–795.