American Journal of Theoretical and Applied Statistics
2022; 11(1): 27-35
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20221101.14
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

On the Coefficient of Determination Ratio for Detecting Influential Outliers in Linear Regression Analysis

Arimiyaw Zakaria[1, *], Benony Kwaku Gordor[2], Bismark Kwao Nkansah[1]

[1]Department of Statistics, University of Cape Coast, Cape Coast, Ghana

[2]Department of Computer Science and Information Systems, Ashesi University, Berekuso, Ghana

Email address:

zakaria.arimiyaw@ucc.edu.gh (A. Zakaria), bgordor@ashesi.edu.gh (B. K. Gordor), bnkansah@ucc.edu.gh (B. K. Nkansah)

[*]Corresponding author

To cite this article:

Arimiyaw Zakaria, Benony Kwaku Gordor, Bismark Kwao Nkansah. On the Coefficient of Determination Ratio for Detecting Influential Outliers in Linear Regression Analysis. *American Journal of Theoretical and Applied Statistics.* Vol. 11, No. 1, 2022, pp. 27-35.
doi: 10.11648/j.ajtas.20221101.14

Received: January 1, 2022; **Accepted**: January 26, 2022; **Published**: February 9, 2022

Abstract: The initial procedure of the Coefficient of Determination Ratio (CDR) for determining outliers in linear regression model is suggested for centred data and declares an observation as an outlier if the CDR value deviates from unity. Although the method performs very well and detects more precisely the requisite outliers than those observed by other well-known detection measures, the cut-off rule approach is a source of subjectivity and the data structure for which the method is designed is also restrictive. In this study therefore, a more rigorous cut-off rule of the same method for identifying influential observations is outlined for an updated method of the CDR that covers the more general case of a non-centred data. A cut-off rule is specified that involves the ratio of quantile values of the Beta distribution. An automated implementation of the procedure is presented that makes use of datasets in the literature and those that are simulated under various conditions of sample size, number and distribution of explanatory variables. The method is now made more generalized in application, objective and reliable as a detection measure than the initial proposal. It therefore provides most appreciable improvement in the explanatory power of linear regression models when the identified outliers are deleted from the data.

Keywords: Outliers, Coefficient of Determination Ratio, Linear Regression, Regression Diagnostics, Influential Observation

1. Introduction

The identification of outliers in linear regression analysis has received attention in many studies (such as those of Chatterjee & Hadi [1, 2]; Hadi [3]; Pena [4]; Barnett & Lewis [5]; Cook & Weisberg [6]; Draper & John [7]; Hadi & Simonoff, [8]; Hawkins [9]; and Lawrence [10]) Most of the outlier detection methods employ case deletion approach, by which the influence of the *i*th observation in the data is measured by computing single-case diagnostics with the *i*th case removed [11-13]. In many case deletion measures, the effect of deleting an observation is determined based on a particular regression result, for instance, the influence of observations on the predicted values, the estimated regression coefficients and its variance-covariance structure. In Zakaria et al. [13], a case deletion measure is proposed as an alternative method for outlier detection based on the coefficient of determination. It is observed that the proposed method in that work, known as the coefficient of determination ratio (CDR), appears more responsive to detecting influential outliers in both simple and multiple linear regression. The method has been compared with some standard case deletion measures and found to be particularly effective in identifying more subtle outlying observations.

Generally, outlier detection measures utilize cut-off rules or threshold values in identifying influential observations in datasets. The work of Zakaria et al. [13] does not provide cutoff values for the CDR measure, but relies on graphical procedure. In that method, if the CDR_i of the *i*th observation

deviates from unity, then the observation is influential. This 'deviation' is a source of subjectivity. The objective of this paper therefore is to determine exact cutoff values for the CDR measure, and provide an automated implementation of the method in R software. The performance of the CDR measure on improving the explanatory power of linear regression models will also be assessed.

In what follows, an update of the theoretical formulation and derivation of the CDR measure is presented. The CDR measure for non-centred linear regression models is also presented. In Section 3, an algorithm for the implementation of the CDR measure is presented. In Section 4, simulation of datasets used in the paper is described. Results and discussion are provided in Section 5, and conclusion follow in the end.

2. Methodology

In the work of Zakaria et al. [13], the CDR measure is proposed based on an assumption that the data under consideration is mean-corrected. In order to prescribe the cut-off value for the measure, we will first provide a review of the CDR to cover the general case of a non-centred data.

2.1. Updating the CDR Measure

Based on the general linear regression model $\mathrm{y} = \mathrm{X}\beta + \varepsilon$, where y is $n\times 1$ vector of responses, X is $n\times(k+1)$ design matrix involving k predictor variables, and corresponding vector of coefficients β, and error term ε, the sum of squares total, SST, may be given as

$$\mathrm{SST} = \mathrm{y}'\mathrm{y} - \left(\frac{1}{n}\right)\mathrm{y}'\mathrm{J}\mathrm{y} \tag{1}$$

and the sum of squares regression, SSR, is also given as

$$\mathrm{SSR} = \hat{\beta}'\mathrm{X}'\mathrm{y} - \left(\frac{1}{n}\right)\mathrm{y}'\mathrm{J}\mathrm{y} \tag{2}$$

where $\mathrm{J} = 11'$, an $n\times n$ matrix of 1s, with $1 = \mathrm{ones}(n, 1)$. From Equations (1) and (2), the coefficient of determination, R^2, is expressed as

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\hat{\beta}'\mathrm{X}'\mathrm{y} - \left(\frac{1}{n}\right)\mathrm{y}'\mathrm{J}\mathrm{y}}{\mathrm{y}'\mathrm{y} - \left(\frac{1}{n}\right)\mathrm{y}'\mathrm{J}\mathrm{y}}$$

The coefficient of determination ratio (CDR) for the ith observation, a measure for detecting influential outliers in linear regression, is given as

$$\mathrm{CDR}_i = \frac{R^2_{(i)}}{R^2},\ i = 1, 2, \ldots, n$$

where $SSR_{(i)}$, and $SST_{(i)}$ are the respective sums of squares regression and total with the ith observation deleted. The CDR expression becomes

$$\mathrm{CDR}_i = \frac{1}{R^2}\times\frac{\mathrm{SSR}_{(i)}}{\mathrm{SST}_{(i)}},\ i = 1, 2, \ldots, n \tag{3}$$

Similar to Equation (2.1), $SST_{(i)}$ may be expressed as

$$SST_{(i)} = \mathrm{y}'_{(i)}\mathrm{y}_{(i)} - \left(\frac{1}{n_{(i)}}\right)\mathrm{y}'_{(i)}\mathrm{J}_{(i)}\mathrm{y}_{(i)},$$

where $\mathrm{y}_{(i)}$ is the corresponding $(n-1)\times 1$ vector after deleting y_i from y. It can be shown [14] that

$$\mathrm{y}'_{(i)}\mathrm{y}_{(i)} = \mathrm{y}'\mathrm{y} - y_i^2.$$

It is noteworthy that $n_{(i)} = n-1$ and

$$\begin{aligned}\mathrm{y}'_{(i)}\mathrm{J}_{(i)}\mathrm{y}_{(i)} &= \left[1'_{(i)}\mathrm{y}_{(i)}\right]^2 \\ &= \left(1'\mathrm{y} - y_i\right)^2 \\ &= \left(1'\mathrm{y} - y_i\right)'\left(1'\mathrm{y} - y_i\right) \\ &= \left(1'\mathrm{y}\right)'\left(1'\mathrm{y}\right) - \left(1'\mathrm{y}\right)y_i - y_i\left(1'\mathrm{y}\right) + y_i^2 \\ &= \mathrm{y}'11'\mathrm{y} - 2\left(1'\mathrm{y}\right)y_i + y_i^2 \\ &= \mathrm{y}'\mathrm{J}\mathrm{y} - 2\left(1'\mathrm{y}\right)y_i + y_i^2\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{SST}_{(i)} &= \mathrm{y}'\mathrm{y} - y_i^2 - \left(\frac{1}{n-1}\right)\left[\mathrm{y}'\mathrm{J}\mathrm{y} - 2\left(1'\mathrm{y}\right)y_i + y_i^2\right] \\ &= \mathrm{y}'\mathrm{y} - y_i^2 - \left(\frac{1}{n-1}\right)\mathrm{y}'\mathrm{J}\mathrm{y} + \left(\frac{2}{n-1}\right)\left(1'\mathrm{y}\right)y_i - \left(\frac{1}{n-1}\right)y_i^2 \\ &= \left(\frac{1}{n-1}\right)\left[n\mathrm{y}'\mathrm{y} - \mathrm{y}'\mathrm{y} - \mathrm{y}'\mathrm{J}\mathrm{y}\right] + \left(\frac{2}{n-1}\right)\left(1'\mathrm{y}\right)y_i - \left(\frac{n}{n-1}\right)y_i^2 \\ &= \left(\frac{1}{n-1}\right)\left[n\left(\mathrm{SST}\right) - \mathrm{y}'\mathrm{y}\right] + \left(\frac{2}{n-1}\right)\left(1'\mathrm{y}\right)y_i - \left(\frac{n}{n-1}\right)y_i^2 \\ &= \left(\frac{1}{n-1}\right)\left[n\left(\mathrm{SST}\right) - \mathrm{y}'\mathrm{y} + 2\left(1'\mathrm{y}\right)y_i - ny_i^2\right]\end{aligned} \tag{4}$$

In line with Equation (2), $SSR_{(i)}$ is stated as

$$\mathrm{SSR}_{(i)} = \hat{\beta}'_{(i)}\mathrm{X}'_{(i)}\mathrm{y}_{(i)} - \left(\frac{1}{n}\right)\mathrm{y}'_{(i)}\mathrm{J}_{(i)}\mathrm{y}_{(i)}$$

where $\mathrm{X}_{(i)} = \mathrm{X} \setminus \mathrm{x}_i$ is the $(n-1)\times(k+1)$ matrix obtained by deleting $\mathrm{x}_i' = (1, x_{i1}, \ldots, x_{ik})$ observation in the ith row of X, and $\hat{\beta}_{(i)}$ is the corresponding vector for $\hat{\beta}$. It can be determined [14] that

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{\hat{\varepsilon}_i}{1-h_{ii}}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_i .$$

The term $\hat{\beta}_{(i)}'\mathrm{X}_{(i)}'\mathrm{y}_{(i)}$ is expanded as

$$\begin{aligned}\hat{\beta}_{(i)}'\mathrm{X}_{(i)}'\mathrm{y}_{(i)} &= \left[\hat{\beta} - \frac{\hat{\varepsilon}_i}{1-h_{ii}}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_i\right]'(\mathrm{X}'\mathrm{y} - \mathrm{x}_i y_i)\\ &= \hat{\beta}'\mathrm{X}'\mathrm{y} - \hat{\beta}'\mathrm{x}_i y_i - \frac{\hat{\varepsilon}_i}{1-h_{ii}}\mathrm{x}_i'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y} + \frac{\hat{\varepsilon}_i}{1-h_{ii}}\mathrm{x}_i'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_i y_i\end{aligned}$$

It is known that, $\hat{y}_i = \mathrm{x}_i\hat{\beta}$, $\hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$, and $h_{ii} = \mathrm{x}_i'(\mathrm{X}'\mathrm{X})^{-1}\mathrm{x}_i$. Thus,

$$\hat{\beta}_{(i)}'\mathrm{X}_{(i)}'\mathrm{y}_{(i)} = \hat{\beta}'\mathrm{X}'\mathrm{y} - \hat{y}_i y_i - \frac{\hat{\varepsilon}_i}{1-h_{ii}}\hat{y}_i + \frac{\hat{\varepsilon}_i}{1-h_{ii}}h_{ii}y_i .$$

Noting that $\hat{y}_i = y_i - \hat{\varepsilon}_i$, we have

$$\begin{aligned}\hat{\beta}_{(i)}'\mathrm{X}_{(i)}'\mathrm{y}_{(i)} &= \hat{\beta}'\mathrm{X}'\mathrm{y} - (y_i - \hat{\varepsilon}_i)y_i - \frac{\hat{\varepsilon}_i}{1-h_{ii}}(y_i - \hat{\varepsilon}_i) + \frac{\hat{\varepsilon}_i}{1-h_{ii}}h_{ii}y_i\\ &= \hat{\beta}'\mathrm{X}'\mathrm{y} - y_i^2 + y_i\hat{\varepsilon}_i - \frac{\hat{\varepsilon}_i}{1-h_{ii}}y_i + \frac{\hat{\varepsilon}_i^2}{1-h_{ii}} + \frac{\hat{\varepsilon}_i}{1-h_{ii}}h_{ii}y_i\\ &= \hat{\beta}'\mathrm{X}'\mathrm{y} - y_i^2 + \frac{\hat{\varepsilon}_i^2}{1-h_{ii}}\end{aligned}$$

The $SSR_{(i)}$ becomes

$$\begin{aligned}SSR_{(i)} &= \hat{\beta}'\mathrm{X}'\mathrm{y} - y_i^2 + \frac{\hat{\varepsilon}_i^2}{1-h_{ii}} - \left(\frac{1}{n-1}\right)\left[\mathrm{y}'\mathrm{J}\mathrm{y} - 2(1'\mathrm{y})y_i + y_i^2\right]\\ &= \left(\frac{1}{n-1}\right)\left(n\hat{\beta}'\mathrm{X}'\mathrm{y} - \hat{\beta}'\mathrm{X}'\mathrm{y} - \mathrm{y}'\mathrm{J}\mathrm{y}\right) + \frac{\hat{\varepsilon}_i^2}{1-h_{ii}} + \left(\frac{2}{n-1}\right)(1'\mathrm{y})y_i - \left(\frac{n}{n-1}\right)y_i^2\\ &= \left(\frac{1}{n-1}\right)\left[n(\mathrm{SSR}) - \hat{\beta}'\mathrm{X}'\mathrm{y}\right] + \frac{\hat{\varepsilon}_i^2}{1-h_{ii}} + \left(\frac{2}{n-1}\right)(1'\mathrm{y})y_i - \left(\frac{n}{n-1}\right)y_i^2\\ &= \left(\frac{1}{n-1}\right)\left[n(\mathrm{SSR}) - \hat{\beta}'\mathrm{X}'\mathrm{y} + \frac{\hat{\varepsilon}_i^2}{1-h_{ii}}(n-1) + 2(1'\mathrm{y})y_i - ny_i^2\right]\end{aligned} \tag{5}$$

Substituting Equations (4) and (5) into Equation (3), the CDR becomes

$$\mathrm{CDR}_i = \frac{n(\mathrm{SSR}) - \hat{\beta}'\mathrm{X}'\mathrm{y} + \dfrac{(n-1)\hat{\varepsilon}_i^2}{1-h_{ii}} + c_i}{R^2\left[n(\mathrm{SST}) - \mathrm{y}'\mathrm{y} + c_i\right]} \tag{6}$$

where $c_i = 2(1'\mathrm{y})y_i - ny_i^2$.

In some applications it is required that the predictor variables in the linear regression model be centred. The centred linear regression model may be stated as

$$y_i = \alpha + \sum_{j=1}^{k}\beta_j(x_{ij} - \bar{x}_j),$$

where, $\alpha = \beta_0 + \sum_{j=1}^{k}\beta_j\bar{x}_j$ and $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n}x_{ij}$. The model is expressed in matrix notation as

$$\mathrm{y} = \alpha\mathrm{j} + \mathrm{X}_c\beta_1 + \varepsilon$$

where, $\beta_1 = (\beta_1, \beta_2, \ldots, \beta_k)'$ and $\mathrm{X}_c = \left(\mathrm{I} - \frac{1}{n}\mathrm{J}\right)\mathrm{X}_1$, which is $n\times k$ matrix, where X_1 is the design matrix without the first

column of 1s.

The CDR_i measure can be computed based on the centred linear regression model. In this regard, the quantities SSR, $\beta'\mathrm{X}'\mathrm{y}$, and h_{ii} are replaced by their counterparts in the centred regression model. The results of the measure based on the centred regression model are the same as those for the non-centred regression model.

The CDR measure is implemented using an algorithm in R. The algorithm takes as input the vector of response variable y, the design matrix X, and the level of significance α. The main output of the algorithm entails the computed values of CDR_i, the cutoff values, and the outliers detected.

2.2. A Cutoff Value for CDR Measure

A relationship between the coefficient of determination, R^2, and the *F*-distribution is given by

$$F = \frac{n-k-1}{k}\frac{R^2}{1-R^2} \tag{7}$$

where, F follows the *F*-distribution, $F_{k,\,n-k-1}$. The probability density function of *F*, $h(f)$, is defined as

$$h(f;k,n-k-1) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{n-k-1}{2}\right)}\left(\frac{k}{n-k-1}\right)^{\frac{k}{2}} f^{\frac{k}{2}-1}\left(1+\frac{k}{n-k-1}\cdot f\right)^{-\frac{1}{2}(n-1)} \tag{8}$$

The probability density function of R^2, $g\left(R^2\right)$, may be obtained by

$$g\left(R^2\right) = h\left(\frac{n-k-1}{k}\cdot\frac{R^2}{1-R^2}\right)\times J_{T^{-1}}\left(F\right), \tag{9}$$

where $J_{T^{-1}}\left(F\right)=\frac{dF}{dR^2}$, the Jacobian of the inverse transformation (T^{-1}) specified in Equation (8). Making substitution into Equation (9) and simplifying gives

$$g\left(R^2\right) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{n-k-1}{2}\right)}\left(R^2\right)^{\frac{k}{2}-1}\left(1-R^2\right)^{\frac{n-k-1}{2}-1} \tag{10}$$

The result shows that R^2 follows the Beta distribution with parameters $\frac{k}{2}$ and $\frac{n-k-1}{2}$.

Ultimately, when all observations in a dataset have equal influence on R^2, the CDR value would be approximately equal to one. When the CDR value markedly exceeds one, it indicates that the corresponding observation is possibly influential. As a result, an observation is identified as an influential outlier if

$$\mathrm{CDR}_i > \frac{B_{\alpha;k/2,(n-k-2)/2}}{B_{\alpha;k/2,(n-k-1)/2}},\ i=1,2,\ldots,n, \tag{11}$$

where, $B_{\alpha;\bullet,\bullet}$ is an upper quantile of the Beta distribution at α level of significance.

3. Simulation of Datasets

Datasets are simulated in R using an algorithm which runs on mvtnorm [15] package. The function is evaluated by specifying different conditions of the regression coefficients, independent variables, and error terms depending on the dynamics of the peculiar dataset as indicated in what follows.

3.1. Dataset 1

The dataset is generated based on a sample size $n=50$. The regression model under consideration is given by

$$y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\varepsilon_i,\ i=1,2,\ldots,n,$$

with the population coefficients generated as $\beta_j \sim N(10,5)$, $j=0,1,2$, and the associated estimates obtained as $\hat{\beta}_{3\times1} \sim N_3\left[\beta,5\left(\mathrm{X}'\mathrm{X}\right)^{-1}\right]$. The independent variables are both sampled from the normal distribution with $X_1 \sim N(30,5)$ and $X_2 \sim N(14,7)$. The mean response vector, $\hat{\mathrm{y}}$, is simulated from $N\left(\mathrm{X}\hat{\beta},5\right)$. The error terms are independent and identically distributed from $N(0,5)$. The observations formed as outliers are 10% of the sample size and are those that lie about three standard deviations (3σ) away from the mean of the distribution. The set of outliers, which are in respect of y values in the dataset is $\{1,11,18,21,37\}$.

3.2. Dataset 2

For this dataset, a sample size of $n=100$ is considered. The linear regression model employed is made up of five predictor variables. The regression parameters are generated as $\beta_j \sim N(20,5)$, $j=0,1,\ldots,5$, and its sample estimate determined as $\hat{\beta}_{6\times1} \sim N_6\left[\beta,5\left(\mathrm{X}'\mathrm{X}\right)^{-1}\right]$. The predictors are simulated as $X_1 \sim N(5,2)$, $X_2 \sim N(14,4)$, $X_3 \sim N(50,10)$, $X_4 \sim N(2,0.05)$, and $X_5 \sim N(70,20)$. The predicted values

of the response variable $\hat{y}$ are generated from $N\left(X\hat{\beta},5\right)$. The error terms are simulated as independent and identically distributed from $N(0,5)$. The observations formed as outliers constitutes 7% of the sample size and lie approximately three standard deviations (3σ) away from the mean of the distribution. The formulated outliers, which are with respect to X values only, is the set $\{5,10,12,26,36,85,93\}$.

3.3. Dataset 3

The dataset is simulated based on a sample size $n=1000$. The linear regression model used consists of ten independent variables. The regression coefficients are generated as $\beta_j \sim N(50,92)$, $j=0,2,\ldots,10$, and its corresponding estimate obtained as $\hat{\beta}_{11\times 1} \sim N_{11}\left[\beta,92\left(X'X\right)^{-1}\right]$. The independent variables are sampled from various distributions: $X_1 \sim N(70,25)$, $X_2 \sim N(20,10)$, $X_3 \sim N(5,0.2)$, $X_4 \sim Poisson(50)$, $X_5 \sim \chi^2_{99}$, $X_6 \sim Poisson(104)$, $X_7 \sim b^{-}(40,0.3)$, $X_8 \sim geom(0.05)$, $X_9 \sim N(210,50)$, and $X_{10} \sim N(257,100)$. The fitted values of the dependent variable $\hat{y}$ are determined from $N\left(X\hat{\beta},92\right)$. The error terms are independent and identically distributed from $N(0,92)$. The observations formed as outliers in the dataset constitutes 1% of the sample size, and lie outside the interval $\mu_j \pm 3\sigma_j$, $j=1,2,\ldots,10$. The formulated outliers is the set $\{1-5,201,202,501,502,503\}$. The subset $\{1,2,3,4,5\}$ are outlying with respect to both y and X values, the subset $\{501,502,503\}$ are outliers in X values alone, and the subset $\{201,202\}$ are outliers in only y values.

3.4. Dataset 4

This data is the artificial data in Table 3 created to illustrate the features of developed package for detecting regression outliers in Siniksaran and Satman [16]. This data is also used in Zakaria et al. [13].

4. Results and Discussion

In this section, the results of the implementation of the automated CDR for the simulated datasets is presented. In each illustration, we show the outliers detected by the CDR_i measure compared with the results from other well-known diagnostics such as the studentised deleted residuals (t_i), the leverage values (h_{ii}), the Cook's distance (D_i), and the difference in fits standardised ($DFFITS_i$). We also assess the performance of the CDR_i measure in improving the R^2 value associated with the underlying linear regression model. The models I, II and III with associated R^2 values obtained from the respective datasets are presented in Table 1. The table shows that all the models are generally significant.

Table 1. *Models extracted from Datasets 1, 2 and 3.*

Parameter	Model					
	I		II		III	
	Estimate	***p*-value**	**Estimate**	***p*-value**	**Estimate**	***p*-value**
β_0	– 67.968	0.311	1585.01	2×10^{-16}	33179.676	2×10^{-16}
β_1	13.279	8.82×10^{-8}	– 3.487	0.56707	13.313	0.146
β_2	8.504	6.21×10^{-7}	11.240	0.00623	– 48.315	0.001383
β_3			1.464	0.63368	– 636.298	3.13×10-9
β_4			– 64.855	0.06697	8.807	0.191980
β_5			13.727	2.83×10^{-10}	30.516	$<2\times10^{-16}$
β_6					14.359	0.002741
β_7					47.323	$<2\times10^{-16}$
β_8					38.137	$<2\times10^{-16}$
β_9					25.713	0.000127
β_{10}					4.952	0.282980
R^2	0.6015		0.4836		0.5031	

4.1. Illustration 1

Table 2 displays the results of the outliers detected for all five detection measures for Dataset 1.

Table 2. *Outliers detected by various measures for Dataset 1.*

Measure	Outliers	No. of Detected Outliers	R^2_{new}	Percentage Change in R^2
CDR_i	1, 11, 18, 37	4	0.9641	60.28
t_i	11, 18, 37	3	0.9155	52.20
h_{ii}	9, 10, 17, 41, 42	5	0.5177	–13.96
D_i	11, 18, 37	3	0.9155	52.20
$DFFITS_i$	11, 18, 37	3	0.9155	52.20
Cut-off value				1.020355

The table shows that the CDR_i measure successfully identifies four of the five outliers in the dataset. Meanwhile, the t_i, D_i and $DFFITS_i$ measures each detect three and the same outliers. For each of these other diagnostics, the outlying observation 1 could not be detected in addition to 21, an observation which is also not detected under CDR. All the actual outliers in the dataset have been completely masked under the h_{ii} measure which detects entirely different set of outliers. Thus, there is deterioration in performance of h_{ii}, an outcome which is not too surprising since it is known to identify high leverage points.

It is noteworthy that when the outlying set of observations detected by the CDR_i measure is removed from the dataset, the estimate of the coefficient of determination (R^2) associated with the linear regression model improves by about 60%. Table 2 shows the percentage increase in R^2 for the five measures. Comparatively, the outliers identified by the CDR_i measure are more influential than those detected by the t_i, D_i and $DFFITS_i$ measures. The set of observations identified by the h_{ii} measure as outliers are actually good leverage values, and hence, their deletion from the dataset would adversely affect the value of the R^2.

4.2. Illustration 2

Table 3 displays the results of outliers detected for all five measures for Dataset 2.

Table 3. *Outliers detected by various measures for Dataset 2.*

Measure	Outliers	No. of Detected Outliers	R^2_{new}	Percentage Change in R^2
CDR_i	5, 10, 12, 25, 26, 36, 59, 74, 85, 93	10	0.9992	106.62
t_i	25, 59, 90	3	0.5076	4.96
h_{ii}	10, 12, 13, 14, 26, 31, 60, 72, 85, 93	10	0.7435	53.74
D_i	5, 10, 12, 26, 59, 74, 85, 93	8	0.8587	77.56
$DFFITS_i$	5, 10, 12, 26, 59, 74, 85, 93	8	0.8587	77.56
Cut-off value				1.009979

From the table, the CDR_i measure successfully detects all the seven outliers in the dataset, but misclassifies three other observations (25, 59 and 74) as outliers. The t_i, could not spot any of the actual outliers, but wrongly identifies three observations as outlying. Two of the real outliers, 5 and 36, have been masked under the leverage values, h_{ii}, which misclassifies half of all observations identified as outliers. Both the Cook's D_i and $DFFITS_i$ diagnostics identify all but one of the actual outliers in the dataset, and misclassify two observations. In this dataset, all the diagnostics are prone to the masking effect, except the CDR_i measure. However, all the measures are susceptible to varying degrees of swamping effect of wrongly identifying observations as outliers with t_i and h_{ii} diagnostics being the worse affected.

It can be observed that there is an overwhelming improvement (106.62%) in the estimated R^2 value when the outliers detected by the CDR_i are deleted from the dataset.

4.3. Illustration 3

Table 4 displays the results of outliers detected for all five measures for Dataset 3. The table shows that the CDR_i measure detects all the real outliers in the dataset, but misclassifies one observation as an outlier. The D_i and $DFFITS_i$ measures also successfully identify all the actual outliers.

Table 4. *Outliers detected by various measures for Dataset 3.*

Measure	Outliers	No. of detected outliers	R^2_{new}	Percentage Change in R^2
CDR_i	1 – 5, 129, 201, 202, 501, 502, 503	11	0.99996	98.76
t_i	227, 305, 741	3	0.49965	– 0.67
h_{ii}	1 – 5, 23, 35, 56, 209, 220, 279, 293, 294, 427, 458, 501, 502, 503, 542, 591, 624, 625, 643, 654, 702, 741, 749, 777, 824, 826, 830, 831, 914, 945	34	0.97353	93.50
D_i	1 – 5, 63, 201, 202, 227, 290, 293, 305, 399, 468, 501, 502, 503, 542, 614, 675, 741, 749, 824, 873	24	0.99995	98.75
$DFFITS_i$	1 – 5, 63, 201, 202, 227, 290, 293, 305, 399, 468, 501, 502, 503, 542, 614, 675, 741, 749, 824, 873	24	0.99995	98.75
Cut-off value				1.000999

However, these measures are vulnerable to swamping effect as they misclassify 50% of the identified observations as outliers. It can be observed that eight of the ten outliers have been identified by the h_{ii} measure, but misclassifies as much as about 76% of what has been detected as outlying observations. Surprisingly, all the actual outliers have been

masked under the t_i diagnostic, but misclassifies three observations as outliers.

The improvements in the value of R^2 are about the same (98.8%) for CDR_i , D_i and $DFFITS_i$ diagnostics though with completely different sets of detected outlying observations. The results show that the outlying observations detected by the CDR_i measure are parsimoniously more precise and influential than the set of outliers detected by any of the other measures.

4.4. Illustration 4

Table 5 displays the results of outliers detected for all five measures for Dataset 4.

Table 5. *Outliers detected by various measures for Dataset 4.*

Measure	Outliers	No. of Detected Outliers	R^2_{new}	Percentage Change in R^2
CDR_i	29, 30, 31, 32	4	0.9872	385.35
t_i	28, 29	2	0.1849	–9.10
h_{ii}	28, 30, 31, 32	4	0.5457	168.29
D_i	28, 29, 30, 31, 32	5	0.9746	379.15
$DFFITS_i$	28, 29, 30, 31, 32	5	0.9746	379.15
Cut-off value				1.032756

It can be observed that the CDR method identifies precisely the four outliers without swamping. These are exactly the same outliers observed by the BCH procedure [17] and the HS procedure [8] out of five procedures in the package of Siniksaran and Satman [16]. In the study of Zakaria et al. [13], the graphical nature of the rule identifies observation 28 in addition to the four realized in this method. The initial proposal [13] of the method could therefore be prone to swamping.

5. Conclusion

An automated implementation of an updated CDR has shown to detect more precisely the requisite outliers than other well-known detection measures in multiple linear regression analysis using datasets in the literature and those simulated under various conditions of sample size, number and distribution of explanatory variables. The performance of the CDR is so realized as it is less prone to masking of actual outliers and swamping of ordinary observations. With a more rigorous inbuilt cut-off rule, the method is now more objective and reliable outlier detection measure than initially proposed.

Notice

Authors are willing to make available upon request the algorithms that are used for the implementations carried out in this work.

Disclosure

The authors report there are no competing interests to declare.

Appendix

Plots of CDR Values for Simulated Datasets.

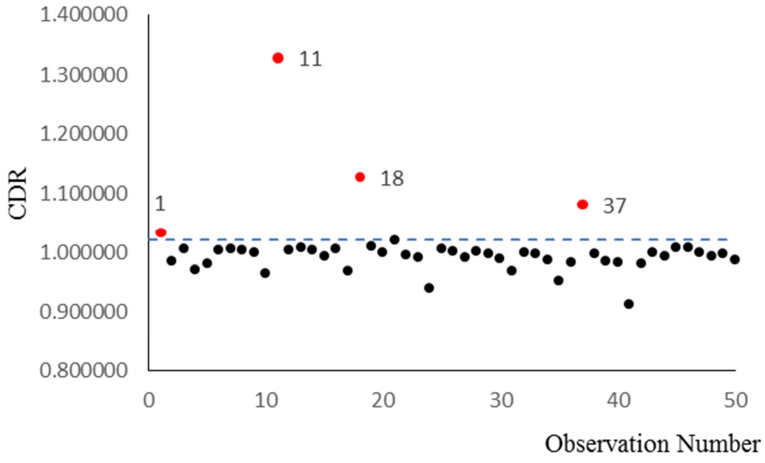

Figure 1. *Plots of CDR values for simulated Dataset 1 showing detected outliers.*

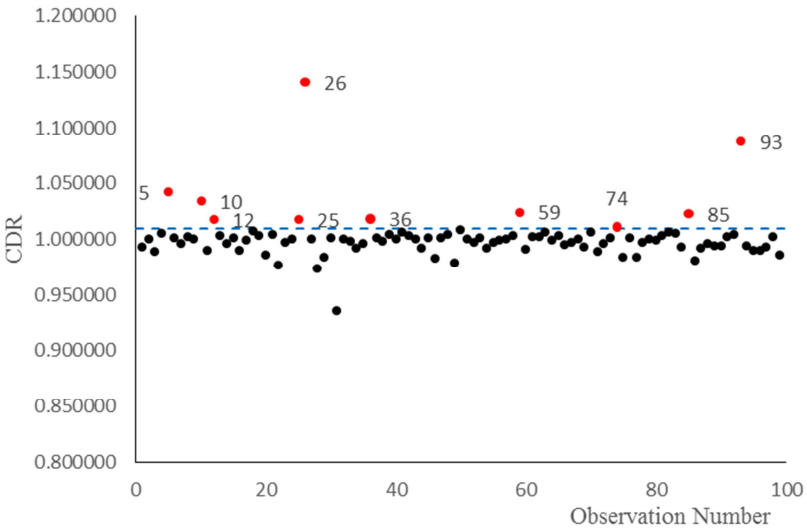

Figure 2. *Plots of CDR values for simulated Dataset 2 showing detected outliers.*

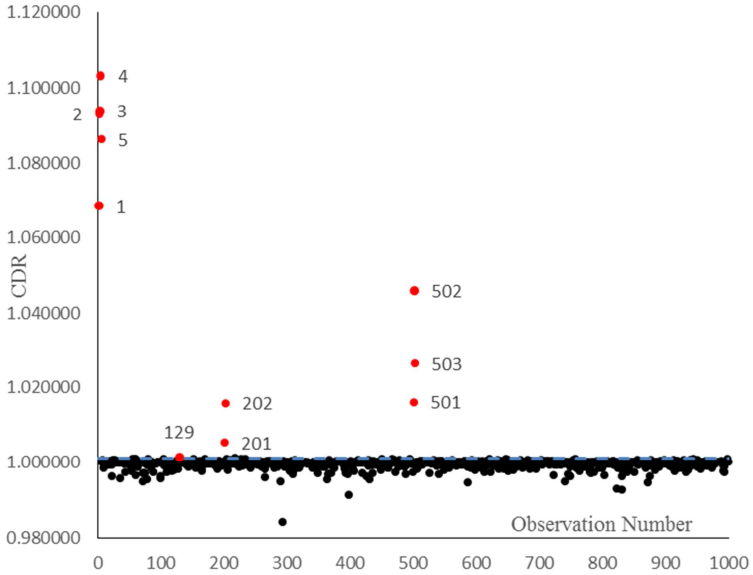

Figure 3. *Plots of CDR values for simulated Dataset 3 showing detected outliers.*

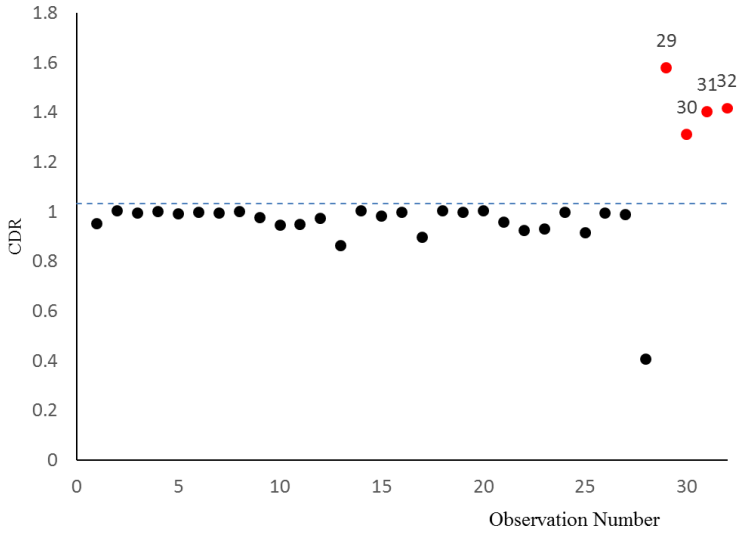

Figure 4. *Plots of CDR values for Dataset 4 showing detected outliers.*

References

[1] Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science, 1* (3), 379-393.

[2] Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example*. New Jersey, NJ: John Wiley & Sons.

[3] Hadi, A. S. (1992). A new measure of overall potential influence in linear regression. *Journal of the Royal Statistical Society, series B (Methodological), 54,* 761-771.

[4] Pena, D. (2005). A new statistic for influence in linear regression. *Technometrics, 47* (1), *1-12*.

[5] Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York, NY: John Wiley and Sons.

[6] Cook, R. D., & Weisberg, S. (1982). *Residuals and Influence in Regression.* New York, NY: Chapman and Hall.

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

[8] Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association, 88* (424), 1264-1272.

[9] Hawkins, D. M. (1991). Diagnostics for the use with regression recursive residuals, *Technometrics*, *33* (2), 221-234.

[10] Lawrence, A. J. (1995). Deletion influence and masking in regression, *Journal of the Royal Statistical Society, Series B (Methodological)*, *57* (1), 181-189.

[11] Cook, R. D. (1977). Detection of influential observations in linear regression, *Technometrics*, *22*: 494–508.

[12] Belsley, D. A., Kuh, E. & Welsch, R. E. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity* (2nd ed.). New Jersey, NJ: John Wiley & Sons.

[13] Zakaria, A., Howard, N. K., & Nkansah, B. K. (2014). On the detection of influential outliers in linear regression analysis, *American Journal of Theoretical and Applied Statistics, 3* (4), 100-106. doi: 10.11648/j.ajtas.20140304.14.

[14] Rencher, A. C. & Schaalje, G. B. (2008). *Linear models in statistics* (2nd ed.). New Jersey, NJ: John Wiley & Sons.

[15] Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2021). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-3, https://CRAN.R-project.org/package=mvtnorm.

[16] Siniksaran, E. & Satman, M. H. (2011). PURO: A package for unmasking regression outliers, *Gazi University Journal of Science*, *24* (1), 59-68.

[17] Billor, N., Chatterjee, S., & Hadi, A. S. (2006). A re-weighted least squares method for robust regression estimation, *American Journal of Mathematical and Management Science*, *26* (3&4), 229-252.