Mathematics and Statistics 11(4): 738-745, 2023
DOI: 10.13189/ms.2023.110416 http://www.hrpub.org

A New Procedure for Multiple Outliers Detection in Linear Regression

Ugah Tobias Ejiofor[1], Arum Kingsley Chinedu[1,*], Charity Uchenna Onwuamaeze[1], Everestus Okafor Ossai[1], Henrrietta Ebele Oranye[1], Nnaemeka Martin Eze[1], Mba Emmanuel Ikechukwu[1], Ifeoma Christy Mba[2], Comfort Njideka Ekene-Okafor[3] Asogwa Oluchukwu Chukwuemeka[4], Nkechi Grace Okoacha[5]

[1] Department of Statistics, Faculty of Physical Sciences, University of Nigeria, Nigeria

[2] Department of Economics, Faculty of Social Sciences, University of Nigeria, Nigeria

[3] Department of Computer Sciences/Mathematics, Faculty of Natural Sciences and Environmental Studies, Godfrey Okoye University, Nigeria

[4]Department of Mathematics and Statistics, Alex Ekwueme Federal University Ndufu Alike, Nigeria

[5] Basic Science Unit, School of Science and Technology, Pan-Atlantic University, Nigeria

Received March 9, 2023; Revised May 20, 2023; Accepted June 11, 2023

Cite This Paper in the following Citation Styles

***(a)**: [1] Ugah Tobias Ejiofor, Arum Kingsley Chinedu, Charity Uchenna Onwuamaeze, Everestus Okafor Ossai, Henrrietta Ebele Oranye, Nnaemeka Martin Eze, Mba Emmanuel Ikechukwu, Ifeoma Christy Mba, Comfort Njideka Ekene-Okafor Asogwa Oluchukwu Chukwuemeka and Nkechi Grace Okoacha, "A New Procedure for Multiple Outliers Detection in Linear Regression," Mathematics and Statistics, Vol.11, No.4, pp. 738-745, 2023. DOI: 10.13189/ms.2023.110416*

***(b)**: Ugah Tobias Ejiofor, Arum Kingsley Chinedu, Charity Uchenna Onwuamaeze, Everestus Okafor Ossai, Henrrietta Ebele Oranye, Nnaemeka Martin Eze, Mba Emmanuel Ikechukwu, Ifeoma Christy Mba, Comfort Njideka Ekene-Okafor Asogwa Oluchukwu Chukwuemeka and Nkechi Grace Okoacha, (2023). A New Procedure for Multiple Outliers Detection in Linear Regression. Mathematics and Statistics, 11(4), 738-745. DOI: 10.13189/ms.2023.110416*

Abstract In this paper, a simple asymptotic test statistic for identifying multiple outliers in linear regression is proposed. Sequential methods of multiple outliers detection test for the presence of a single outlier each time the procedure is applied. That is, the most severe or extreme outlying observation (the observation with the largest absolute internally studentized residual from the original fit of the mode to the entire observations) is tested first. If the test detects this observation as an outlier, then this observation is deleted, and the model is refitted to the remaining (reduced) observations. Then the observation with the next largest absolute internally studentized residual from the reduced sample is tested, and so on. This procedure of deleting observations and recomputing studentized residuals is continued until the null hypothesis of no outliers fails to be rejected. However, in this work our method or procedure entails calculating and uses only one set of internally studentized residuals obtained from fitting the model to the original data throughout the test exercise, and hence the procedure of deleting an observation, refitting the data to the remaining observations (reduced values) and recomputing the absolute internally studentized residuals at each stage of the test is avoided. The test statistic is incorporated into a technique (procedure) that entails a sequential application of a function of the internally studentized residuals. The procedure is a straightforward multistage method and is based on a result giving large sample properties of the internally studentized residuals. Approximate critical values of this test statistic are obtained based on approximations that depend on the application of the Bonferroni inequality since their exact values are not available. The new test statistic is very simple to compute, efficient and effective in large data sets, where more complex methods are difficult to apply because of their enormous computational demands or requirements. The results of the simulation study and numerical examples clearly show that the proposed test statistic is very successful in the identification of outlying observations.

Keywords Asymptotic, Bonferroni Upper Bound, Critical Values, Internally Studentized Residuals, Outlier, Robust, Regression Diagnostics, Test Statistic

1 Introduction

Data collected by researchers commonly contain one or few unusual observations (outliers) that do not seem to belong to the pattern of variability exhibited by other observations in the set. In a regression problem, observations are said be outlying based on how unsuccessful the fitted model is in accommodating them. Observations corresponding to excessively large residuals (the difference between the observed value and the value obtained from fitting a model to the data) are usually treated as outliers. A book by Barnett and Lewis [1] is a very valuable comprehensive book on the treatment of outliers.

Outliers may occur as a result of gross mistakes or errors during the collection, recording, or transcription of the data or as a result of gross deviation from prescribed experimental procedures. In a least squares analysis of linear regression models , outliers can be excessively influential in the estimation of parameters. The presence of outliers in a data set can mar or distort the inferential process such as estimation, and hence render the standard tests of hypotheses meaningless. Their presence, if unnoticed and not tackled properly, can grossly affect standard errors of estimators (by inflating standard error of estimators), reducing the power of test statistics, afflicting confidence intervals, and seriously distorting conclusions about relationships between variables. The existence of outliers can severely distort the summary quantities and analyses of data.

Of particular importance or concern to data analysts in many fields is the detection of outliers in linear regression modeling, because of the widespread application or use of regression technique. Also, outliers have been a serious concern in the analysis of linear regression models basically because of the vulnerability shown by regression methods or techniques in the presence of outliers. Excellent book-length treatments of outliers include [1], [2], [3], [4]. See also papers of [5],[6], [7],[8], [9], [10], [11], [12], [13], [14].

Screening data collected by field workers for outliers is an integral part of model building and this has drawn a great deal of attention in regression diagnostics, particularly on the identification of a single outlier. Attention has now shifted to the more tasking and computationally tedious or burdensome problem of detecting multiple outliers, which is the main focus of this work.

There are two main types of methods that have been used to identify multiple outliers. They are the block methods and the stepwise methods. Stepwise procedure is also called "consecutive" or "sequential" method. The steppwise procedure is used to test for the presence of a single outlier each time the procedure is applied. There are two applications of the stepwise method. They are the 'Inward' and "Outward" applications. In applying an inward stepwise procedure, the most sever outlier is tested first using the entire n observations first, then the next most sever, and so on. The procedure terminates when the kth test does not detect an outlier. The number of outlyng observations declared is $k-1$. This procedure has the advantage that the number of most likely observation in the data set need not be specified in advance. Outward stepwise method , on the other hand, requires the prespecification of k, the number of most likely observations in the data set. Unlike inward stepwise procedure, the outward procedure begins by testing the least sever outlying observation among the k most outlying observations. If the test detects or identifies this least sever outlying observation as an outlier, all other more extreme outlying observations in the prespecied k most outlying observations most outlying observation are judged to be outliers as well, and the test procedure terminates. Otherwise, the least extreme outlying is deleted from the set (k is reduced by one) and the test is carried out on the least extreme outlier outlying observation in the remaining group of $k-1$ suspected outliers, and so on. Demerits of this procedure are (i) the need to prespecify k . (ii) the enormous computational effort involved, and (iii) Swamping may occur if k is too large.

Block methods use a set of $k > 1$ observations at each phase or stage of the test and consider the outlyingness of the group as a whole. The procedure requires the user to specify k, the maximum number of suspected outlying observations believed to be in the data set, before the procedure can be applied. They entail grouping the data into a clean subset without outlying obsrvations and a contaminated subset that consists all the potential outlying obsrvations and then test the outlyingness of the complementary subset. The block procedures are liable or prone to swamping and are computationally intensive.

The earliest method for multiple outliers identification in regression models is credited to Mickey at al.[15]. Mickey at al.[15] proposed a method of multiple outliers identification in the context of a stepwise-regression calculation. Mickey at al.[15] applied a stepwise regression approach and included dummy variables that identify outliers to the basic model. The method first finds the single outlying observation whose removal causes the highest reduction in the sum of squared residuals, then the next observation outlying observation whose removal further engenders reduction in the sum of squared residuals, and so on. These observations are then ordered accordingly. They maintained that cases can be considered as outlying observation in a regression structure if their deletion results in substantial reduction of the residual sum of squares. They used stepwise regression programs to implement this procedure.

Considering multiple outliers, Gentleman and Wilk [16] adopted a method for group outliers detection that involves identifying the "k most likely subset of outliers" in a data set, the removal or deletion of which causes the largest reduction in the residual sum of squares. That is, the method was aimed at identifying the most outlying subset of k observations whose removal or deletion produced the highest reduction in the sum of squared residuals. Setbacks of their procedure are (a) the problem of specifying k in advance and (b)the tremendous or enormous computational efforts are involved. They presented the reduction in the sum of squared residuals resulting from refitting the model after deleting k observations $i_1, i_2, ..., i_k$ as

$$Q_k = \sum_{i}^{k} t_{i_r}^2$$

and based the test statistic for deleting the k most likely outliers on the maximum of the $\binom{n}{k}$ possible values of Q_k. However, exact critical values of this test statistic are not available, and approximations are obtained via the Bonferroni inequality. It

has been pointed out that this method is not well suited for routine application.

Marasinghe [17] proposed a new test statistic, F_k, for flagging multiple outlying observations in linear regression. Initially, a subset consisting of k observation to be tested is selected as follows: The first observation in the subset is the observation that has the largest absolute studentized residual from the original fit of the model. This value or observation is then deleted , and the model is refitted using the remaining $n-1$ observations. The observation that has the largest absolute studentized residual from this refitting is then included in the subset. This process is continued until the subset consisting of k observations is determined. The test statistic F_k is defined as

$$F_k = \frac{(S - Q_k^*)}{S}$$

where $S = (n-p)\hat{\sigma}$, $Q_k^* = \sum_i^k a_{i_r}^2$ and $a_i = \frac{e_i}{\sqrt{1-h_{ii}}}$ is a transformation of the studeiitized residuais termed "adjusted residuals". Using F_k, one rejects the null hypothesis of no-outliers when F_k is smaller than a specified critical value. If the test statistic F_k is found to be significant, the most extreme outlying observation in the subset as determined by the largest studentized residual is deleted and the procedure is repeated for the $(k-1)$ observations in the subset using the remaining $n-1)$ observations. The procedure is terminated when a test fails to reject the null hypothesis of no-outliers. The procedure proposed by Marasinghe [17] is a multistage procedure for flagging multiple outliers in linear regression models.

The multistage method is effective in some cases, but demands prespecifying the maximum number of outliers (in this case k) one can then detect. The method can suffer gravely if the chosen value fo k is either larger or smaller than the actual number of outliers present in the data. It entails a considerable amount of computing time and is computationally demanding.

In this work, a straightforward stage-wise procedure for identifying of multiple outliers in the response variable in least squares analysis of linear regression is proposed. Unlike the aforementioned works, the procedure in this paper does not involve prespecifying k or deletion (removal) of observation of any kind throughout the procedure. The test statistic uses a function of the internally studentized residuals to sequentially detect a set of outlying observations. This method requires computing only one set of internally studentized residuals from fitting the model to the original data. It involves the use of internally studentized residuals computed from the initial (original) fit of the model to the entire data, thereby circumventing the cumbersomeness and tedium associated with the technique of refitting the model to the reduced data at each stage the null hypothesis is being tested. Also, the problems associated with specifying or stipulating in advance the number of outliers to be deleted are also completely avoided. The only input needed is the initial internally studentized residuals computed from fitting regression model to the entire data. Further advantages of this method are saved in the computing time and ease of computation. Above all, it is well suited for routine application in applied regression analysis and can easily be applied by the researcher.

One of the main methods for outlier detection is the analysis residuals. Suppose we have a classical linear regression model

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

where $\mathbf{Y}$ is the $n \times 1$ vector of observations, $\boldsymbol{X}$ is an $n \times$ p matrix of constants, $\boldsymbol{\beta}$ is a p $\times 1$ vector of unknown parameters to be estimated and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of normally distributed errors. Assuming that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, the least squares estimator of $\boldsymbol{\beta}$ in (1) is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

and vector of residuals is

$$\begin{aligned}\mathbf{e} =& \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ =& (\boldsymbol{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\varepsilon}.\end{aligned}$$

The ith element of $\mathbf{e}$ is $e_i = y_i - \hat{y}_i$, where $\hat{y}_i$ is the predicted value of y_i. The variance-covariance matrix of $\mathbf{e}$ is

$$\text{Var}(\mathbf{e}) = (\boldsymbol{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\sigma^2.$$

Using $\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-p}$ as an estimate of σ^2, the estimated variance-covariance matrix of $\mathbf{e}$ becomes

$$\widehat{Var}(\mathbf{e}) = \left(\boldsymbol{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'\right)\hat{\sigma}^2. \tag{2}$$

The estimated variance of the ith residual e_i is

$$\widehat{Var}(e_i) = (1 - h_{ii})\hat{\sigma}^2, \tag{3}$$

where h_{ii} is the ith diagonal element of matrix $\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'$, called the hat matrix and $(1 - h_{ii})\hat{\sigma}^2$ is the ith diagonal element of $\widehat{Var}(\mathbf{e})$.

Residuals from the least squares analysis of a general linear regression model play a key role in regression diagnostics. They are used in various graphical plots (such as the rankit-plots) and procedures for checking the adequacy of the model. Numerous graphical plots and numerical techniques for checking model assumptions and adequacy using residuals are ubiquitous (or abundant) in the literature. The ordinary least squares residuals e_i, however, have certain deficient features that dwarf their useful roles in regression diagnostics. It is known that the ordinary least squares residuals e_i are not independent or homoscedastic, and their joint distribution depends on $\boldsymbol{X}$ through $\boldsymbol{H}$. These deficiencies restrict their usefulness in regression diagnostics.

Therefore, a transformed (standardized) version of them that is free of these nuisance quantities is preferable. For use in diagnostic purposes, several standardizations or transformations of the ordinary residuals have been proposed to overcome their deficiencies (see [4]). Prominent among them is the internally studentized residual which has a representation of the form

$$R_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \tag{4}$$

The internally studentized residuals have sound statistical properties that make them versatile in regression diagnostics. The Studentized residuals are used as replacements for the ordinary residuals in graphical procedures and in other regression diagnostics. Apart rom being diagnostic tools, they are as well important parts of other regression diagnostic measures.

The internally studentized residuals have been studied by many authors (see [18, 19]). Define ξ_i=$\frac{R_i}{\sqrt{n-p}}$. Ellenberg [19] derived ξ_i and showed that their joint distribution is a multivariate Inverted Students function. Ellenberg [19] showed that the marginal probability density function for any ξ_i is a univariate Inverted-Students Function. Díaz-García and Gutiérrez-Jáimez [18] proposed a straightforward alternative method to the the work by Ellenberg [19]. These studies by [18], [19] were motivated by a concern about outliers. The results of Ellenberg [19] have been used by many authors in deriving test statistics for outliers. Lund [20] used the result of [19] and obtained approximate critical values of the maximum studentized residual statistic.

2 The Proposed Test Statistic

A good number of applications in the physical, medical and social sciences use regression models. Many regression problems require the analysis of large datasets. Franklin and Hariharan [21] pointed out that the internally studentized residuals follow the standard normal distribution as the sample size n increases. Chatterjee and Hadi [4] said that the internally studentized residuals should approximately have a standard normal distribution, especially when the sample size is large and that the lack of independence assumption concerning the the internally studentized residuals may be ignored. Most graphic techniques for regression diagnostics hinge on this assumption. The assumption that the internally studentized residuals are (approximately) equivalent to a random sample from a standard normal distribution is used in regression diagnostics to check the validity of some of the model assumptions (see [4],[21]). For instance, the rankit-plot of the ordered internally studentized residuals is one of the popular regression diagnostic techniques that makes use of this assumption (see [4],[21]). Based on these asymptotic properties of these residuals, we propose a new asymptotic test statistic for multiple outliers identification. Define

$$T_i = R_i^2,\ i = 1, 2, 3, ...n. \tag{5}$$

It can be shown that asymptotically T_i has a Chisquare distribution with $\nu = 1$ degree of freedom :

$$T_i \simeq \chi^2_{(\nu=1)}. \tag{6}$$

Consider a set T defined by

$$T = \{T_1, T_2, T_3, ..., T_n\}. \tag{7}$$

Let N be the number of subsets each of size $n\prime$ that can be formed using the elements in T, meaning that we have a total of N subsets ($N = \binom{n}{n\prime}$) each of size $n\prime$. Also let

$$Z_j,\ j = 1, 2, 3, , ..., N \tag{8}$$

be the sum of the $n\prime$ elements of the jth subset. Then, from the knowledge of distribution theory, it can be shown that Z_j has a Chi-square distribution $f_{Z_j}(z)$ with sum of squared residuals degrees of freedom. That is,

$$Z_j \simeq \chi^2_{(n\prime)}. \tag{9}$$

We propose

$$Z_{n\prime} = max\{Z_1, Z_2, ..., Z_N\} \tag{10}$$

as a statistic for flagging the subset with the $n\prime$ most likely outlying observations based on the maximum of the ($N = \binom{n}{n\prime}$) values of $Z_{n\prime}$. Using $Z_{n\prime}$, one rejects the null hypothesis of no-outliers when $Z_{n\prime}$ exceeds a specified critical value.

2.1 Evaluation of critical values of $Z_{n\prime}$

In this section we describe a method of calculating critical values for the proposed test statistic for large sample sizes. Exact critical values of $Z_{n\prime}$ are not available, and hence approximations of them can be obtained via the Bonferroni inequality, denote by $Z_{n\prime}(\alpha)$ the approximate critical value of $Z_{n\prime}$. Finding $Z_{n\prime}(\alpha)$ would provide a standard and an objective way of using $Z_{n\prime}$ to identify the subset with the $n\prime$ most outlying observation(s). This idea motivates the current work. To obtain $Z_{n\prime}(\alpha)$, we need to evaluate the equation

$$\int_{Z_{n\prime}(\alpha)}^{\infty} n\prime f_{Z_j}(z)\mathrm{dz} = \alpha. \tag{11}$$

numerically. The computed approximate critical values $Z_{n\prime}(\alpha)$ for significance levels α=0.01 are 0.05 and $\nu = 1, 2, 3, 4, 5, 6$ and 7 are presented them in Table 1 below. They were obtained using the Mathematica software version 12.

Table 1. Upper bounds for the Critical values of $Z_{n\prime}$ for multiple outliers identification in linear regression.

Subset Size (n$\prime$)	$\alpha = 0.01$	$\alpha = 0.05$
1	6.6349	3.8415
2	10.5966	7.3778
3	13.7064	10.2355
4	16.4239	12.7619
5	18.9074	15.0863
6	21.2318	17.2722
7	23.4398	19.3588

The test statistic $Z_{n\prime}$ is used to test a no-outliers hypothesis at each stage. For each value of $n\prime = 1, 2, 3, ..., N$ (starting with $n\prime = 1$), the value of $Z_{n\prime}$ is computed and compared with the corresponding critical value $Z_{n\prime}(\alpha)$. If $Z_{n\prime}$ is larger than the critical value $Z_{n\prime}(\alpha)$, null no-outliers hypothesis is rejected and the observation(s) in that subset are judged to be outliers. The procedure is stopped when a test fails to reject the null hypothesis of no-outliers.

3 The test procedure

With respect to the model discussed in the introduction, the test procedure commences with obtaining a set consisting of the squares of the internally studentized residuals (see equation (5)). To screen the response variable Y for a suspected set of outliers, we begin by forming subsets ($N = \binom{n}{n\prime}$) of the elements of (5). Usually we start with $n\prime = 1$ to obtain n subsets each consisting of one value. Then we compare the largest value among these subsets with the critical value, $Z_1(\alpha)$ and make a decision. If $Z_1 > Z_1(\alpha)$, we reject the no-outlier null hypothesis and proceed to the second stage. Next, we put $n\prime = 2$ and obtain $N = \binom{n}{2}$ subset each consisting of two elements. We compute the sum of the values in each subset and compare the largest sum with the critical value $Z_2(\alpha)$. If $Z_2 > Z_2(\alpha)$, we reject the no-outlier null hypothesis and proceed to the third stage. This process is continued (with the same internally studentized residuals from the initial fit of the model) until the null hypothesis of no-outliers fails to be rejected, thus determining thee subset with the most outlying observations) present.

4 Application to real data

We now demonstrate the use of the new test statistic in flagging the presence of multiple outliers in linear regression. We demonstrate this using real data which are very frequently referred to in regression diagnostics.

4.1 Stack loss data

Here we present an example that is based on stack loss data from the work of [22], which has been used by many authors in regression diagnostics, particularly for multiple outliers. The stack loss dataset has 21 observations with 3 independent variables (see Table 2). The data set consists of the dependent variable (Y) which is the percent of the in-going ammonia that is lost by escaping in the unabsorbed nitric oxides. Independent variables are as follows: X_1 =air flow (which reflects the plant's operation rate), X_2 =temperature of the cooling water in the coils of the absorbing tower for the nitric oxides, X_3 =concentration of nitric acid in the absorbing liquid.

Column 6 in Table 2 contains the values of the set T (see equation (7)). For a fixed $n\prime$, we obtain $N = \binom{n}{n\prime}$ subsets of the values in column 6. Among these subsets, we pick the subset whose elements give the largest sum. It is this sum (the value of $Z_{n\prime}$) that we compare with the critical value $Z_{n\prime}(\alpha)$. We begin with smallest value of $n\prime$ (in this case $n\prime = 1$) and fix $\alpha = 0.05$ for all $n\prime$ to be used in this example. The number of outliers to be flagged or declared is determined when the value of test statistic exceeds its critical value at a given stage.

(1). For $Z_{n\prime} = 1$, we have only $\binom{n}{1} = 21$ subsets each containing one element. The subset with the largest element is $\{6.9602\}$. From Table 1, $Z_1(0.05) = 3.8415$. So we declare the value of Y with serial number 21 to be an outlier, since $Z_1 = 6.9602 > Z_1(0.05) = 3.8415$.

Table 2. Stack-Loss Data

sn	Air flow (X_1)	Cooling water (X_2)	Acid cont.(X_3)	Stack loss (Y)	$T_i = R_i^2$
1	80	27	89	42	1.4241
2	80	27	88	37	0.5124
3	75	25	90	37	2.3902
4	62	24	87	28	3.5412
5	62	22	87	18	0.2939
6	62	23	87	18	0.9317
7	62	24	93	19	0.6952
8	62	24	93	20	0.2351
9	58	23	87	15	1.0931
10	58	18	80	14	0.1908
11	58	18	89	14	0.7819
12	58	17	88	13	0.9381
13	58	18	82	11	0.2303
14	58	19	93	12	0.0003
15	50	18	89	8	0.6548
16	50	18	86	7	0.0896
17	50	19	72	8	0.3736
18	50	19	79	8	0.0235
19	50	20	80	9	0.0412
20	56	20	82	15	0.2061
21	70	20	91	15	6.9602

(2). For $n\prime = 2$, we obtained $\binom{n}{2} = 210$ subsets each with two elements. Out of these 210 subsets, the subset $\{6.9602, 3.5412\}$ has the largest sum value of 10.5014 ($Z_2 = 10.5014$). From Table 1, $Z_2(0.05) = 7.3778$. So we declare values of Y with serial numbers 21 and 4 to be outliers, since $Z_2 = 10.5014 > Z_2(0.05) = 7.3778$.

(3). For $n\prime = 3$, we obtained $\binom{n}{3} = 1330$ subsets each with three elements. Out of these 1330 subsets, the subset $\{6.9602, 3.5412, 2.39018\}$ has the largest sum of 12.8916 ($Z_3 = 12.8916$). From Table 1, $Z_3(0.05) = 10.2355$. So we declare values of Y with serial numbers 21, 4 and 3 to be outliers since $Z_3 = 12.8916 > Z_3(0.05) = 10.2355$.

(4). For $n\prime = 4$, we obtained $\binom{n}{4} = 5985$ subsets each with four elements. Out of these 5985 subsets, the subset $\{6.9602, 3.5412, 2.39018, 1.42406\}$ has the largest sum of 14.3157 ($Z_4 = 14.3157$). From Table 1, $Z_4(0.05) = 12.7619$. So we declare values of Y with serial numbers 21, 4 , 3 and 1 to be outliers since $Z_4 = 14.3157 > Z_4(0.05) = 12.7619$.

(5). For $n\prime = 5$, we obtained $\binom{n}{5} = 20349$ subsets each with five elements. Out of these 20349 subsets, the subset $\{6.9602, 3.5412, 2.39018, 1.42406, 1.09313\}$ has the largest sum of 15.4088 ($Z_5 = 15.4088$). From Table 1, $X_5(0.05) = 15.0863$. So we declare values of Y with serial numbers 21, 4 , 3, 1 and 9 to be outliers since $Z_5 = 15.4088 > Z_5(0.05) = 15.0863$.

(6). For $n\prime = 6$, we obtained $\binom{n}{6} = 54264$ subsets each with six elements. Out of these 54264 subsets, the subset $\{3.5412, 6.9602, 2.39018, 1.42406, 1.09313, 0.93814\}$ has the largest sum of 16.3469 ($Z_6 = 16.34698$). From Table 1, $Z_n(0.05) = 17.3588$ and $16.3469 < 17.3588$. Further computations for $n\prime = 7, 8, .., 21$ did not yield any significant results

(failed to reject the no-outlier null hypothesis). Table 3 gives a summary of the results for each phase or stage of the multistage search for a subset of $n\prime$ most likely outlying observations.

Table 3. Diagnostic measure for the Stack-loss data

(n$\prime$)	$\binom{21}{n\prime}$	Subset	$Z_{n\prime}$	$Z_{n\prime}(0.05)$
1	21	{6.9602}	6.9602	3.8415
2	210	{6.9602, 3.5412}	10.5014	7.3523
3	1330	{6.9602, 3.5412, 2.39018}	12.8917	10.1984
4	5985	{6.9602, 3.5412, 2.39018, 1.42406}	14.3157	12.717
5	20349	{6.9602, 3.5412, 2.39018, 1.42406, 1.09313}	15.4088	15.0368
6	54264	{6.9602, 3.5412, 2.39018, 1.42406, 1.09313, 0.93814}	16.3469	17.2186
7	116280	{6.9602, 3.5412, 2.39018, 1.42406, 1.09313, 0.93814, 0.6548}	17.2787	19.2989

4.2 Discussion

As aforementioned, the "stack loss" dataset is well-known and is one of the most referred datasets in the context of multiple outliers identification in regression diagnostics. In a well-researched article, Dodge [23] investigated the history of this dataset and pointed out that about 26 distinct sets of detected outliers have been found by various methods used in analyzing this data set. According to him, the most cited set being a set containing the observations 1, 3, 4, and 21. These data have been investigated and analyzed by many authors and discussed extensively in the books by [23], [24], [25], as well as in the papers by [11], [18].

The Stack loss data set was analyzed by Atkinson [5]. One set of conclusions given by Atkinson [5] is that observations (values of the response variable Y) with serial numbers 1, 3, 4 and 21 are outliers. Note that the stack loss data set contains three covariates and this conclusion by Atkinson [5] is given when all three covariates are used in analysis.

Nurunnabi et al. [6] introduced a robust influence distance that can identify multiple influential observations (IOs) which are named ID, and proposed a sixfold plotting technique that is designed for the identification and classification of multiple outliers, high leverage points and influential observation on the same graph at one time (simultaneously) in linear regression. Their method correctly classifies the observations 1, 3, 4, 9 and 21 as outliers. They also applied the standardized Studentized residual statistics and the Standardized LMS residuals. The standardized Studentized residual statistic identify only one observation (case 21) as an outlier but mask all other potential outliers. The Standardized LMS residuals identify five cases 1,2, 3, 4, 9 and 21 as outliers.

Imon [26] developed a generalized version of DFFITS based on group deletion and then proposed a new technique to identify multiple influential observations. The GDFFIT was applied to the Stack Loss Data and it correctly identifies all five influential cases, namely1, 2, 3, 4 and 21 with case 2 as an influential observation and cases 1, 3, 4 and 21 as outliers. He also applied the reweighted least squares (RLS) introduced by [21]. The robust RLS technique identifies cases 1, 3, 4 and 21 as outliers. By applying the multistage procedure, Marasinghe [17] declared observations 21 and 4 as the only outliers. In summary, we consider observations of the response variable Y with serial numbers 1, 3, 4, 9 and 21 as outliers.

5 Simulation Study

In this section we conduct a simulation study using using R software to evaluate the performance of the test statistic proposed in Section 2. We used a simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ in the simulation study. The residuals are not in any way affected by the particular values of β_0 and β_1 used in the simulation. We therefore set $\beta_0 = 1$ and $\beta_1 = 2$. Thus, $Y = 1 + 2X + \varepsilon$.

Clean values of Y were obtained as follows: (a) n values of X were sampled from a normal distribution with mean $\mu = 2$ variance $\sigma^2 = 0.11$ (both arbitrarily chosen).(b) One set of $\hat{Y}$ values was generated by adding the n values of X to $Y = 1 + 2x$. (c) Then 1000 sets of ε values were sampled from a standard normal distribution and each added to $\hat{Y}$ to generate 1000 sets of Y values. The sample sizes are taken as $n = 10, 15, 20$ and 25. At each value of n, we simulated 1000 sets of Y values.

Then following the introduction of outliers, we introduced (planted) outliers using the formula below:

$$y^*_{ij} = \lambda * \max(Y_j) + y_{ij}, j = 1, 2, 3, , ..., 1000, i = 1, 2, 3, , ..., n. \quad (12)$$

where y_{ij} is the selected value of Y_j to be contaminated or polluted, y^*_{ij} is the polluted y_{ij}, $\max(Y_j)$ is the maximum of the observations in the vector Y_j, and λ is the magnitude of outlier, see ([27], [28], [29], [30], [31]). Up to four outliers per sample were planted. We denoted the number of outliers per sample by η. The table entries are percentage of planted outliers that were correctly identified by the proposed test statistic. A nominal size of $\alpha = 0.05$ was used throughout the test. For clarity of purpose, a schematic notation for the simulation study is shown in Table 4.

Table 5 shows the percentage of correctly identified outlier by the proposed test statistic. The behavior of the test statistic under a particular set of λ and n values was examined by varying the magnitude λ of the outlier and the sample size n throughout the simulation study. Perusing through the contents of Table 5 shows clearly the effectiveness of the proposed test statistic in detecting outliers. The results of the simulation study show that the statistic can flag or identify a reasonable number of outlying observations for each set of λ and n.

Table 4. Schematic notation for the simulation study for a sample of size n

Predictor	Response Variables			
X	Y_1	Y_2	$\cdots$	Y_{1000}
x_1	y_{11}	y_{12}	$\cdots$	y_{11000}
x_2	y_{21}	y_{22}	$\cdots$	y_{21000}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_{n1}	y_{n2}	$\vdots$	y_{n1000}
y^*_{ij}	$y^*_{i1} = m * \max(Y_1) + y_{i1}$	$y^*_{i2} = m * \max(Y_2) + y_{i2}$	$\cdots$	$y^*_{i1000} = m * \max(Y_{1000}) + y_{i1000}$

6 Conclusions

The method introduced in this work entails a sequential application of a function of the internally studentized residuals obtained to detect or flag outliers. The proposed method uses

Table 5. Percentage of outliers correctly identified

λ	η	Sample Sizes			
		10	15	20	25
1.5	2	47	58	61	72
2.0	2	59	67	73	82
1.5	3	61	75	79	88
2.0	3	64	78	89	96
1.5	4	66	77	83	91
2.0	4	70	79	89	97

only initial internally studentized residuals computed from fitting the model to the original entire data throughout the search for outliers in the data. Thus, the tediousness associated with the method of refitting the model to the reduced data and recomputing the internally studentized residuals at each stage of the search for outliers is completely eschewed. It is very simple to use and with a comprehensible interpretation, that can be a valuable tool in applied statistical analysis. The proposed new method is very applicable in large datasets in high dimension, where more complex procedures are difficult to use because of their high computational demands. An application to a real data set shows that the method proposed herein is equally good as those of other authors. Also, the simulation study conducted shows that the proposed test statistic is effective enough in detecting outlying observations. In summary, the approach is simple and its advantages are saving in computing time and ease of application.

REFERENCES

[1] Barnett V., Lewis T., "Introduction," Outliers in Statistical Data, 2nd ed, John Wiley and Sons, 1978, pp. 6-15.

[2] Fox J., "Outlying and influential Data," Regression diagnostics: an introduction, Sage, Newbury Park, 1991.

[3] Rousseeuw P. J., Leroy A. M., "Introduction," Robust regression and outlier detection, Wiley, New York, 1987.

[4] Chatterjee S., Had A . S., "Regression Diagnostics: Detection of Model Violations," Regression analysis by example, Wiley, New York, 2015.

[5] Atkinson A.C., "Fast Very Robust Methods for the Detection of Multiple Outliers," Journal of the American Statistical Association, vol. 89, no. 428, pp. 1329–1339, 1994. DOI: 10.1080/01621459.1994.10476872.

[6] Nurunnabi A.A.M., Nasser M., Imon A.H.M.R., "Identification and classification of multiple outliers, high leverage points and influential observations in linear regression," Journal of Applied Statistics. vol. 43, no. 3 , pp. . 509-525, 2016. DOI: 10.1080/02664763.2015.1070806.

[7] Hadi A.S., "A new measure of overall potential influence in linear regression," , Computional Statistics and Data Analysis, vol. 14, no. 1, pp. 1–27, 1992. DOI: 10.1016/0167-9473(92)90078-t.

[8] Hadi A.S., Simonoff J.S., "Procedures for the identification of multiple outliers in linear models," Journal of the American Statistical Association, vol. 88, no. 424 ,pp. 1264–1272 1993. DOI:10.2307/2291266.

[9] Bagdonavicius V., Petkevicius L., " A new multiple outliers identification method in linear regression," Metrika, 2019. .DOI:10.1007/s00184-019-00731-8

[10] Peña D., "A new statistic for influence in linear regression, Technometrics, vol.47, no. 1, pp. 1–12, 2005. DOI:10.1198/004017004000000662.

[11] Denby L., Mallows C.L., " Two diagnostic displays for robust regression analysis," Technometrics, vol. 19, no.1 pp. 1-13, pp. 1977. DOI:10.2307/1268248.

[12] Rousseeuw P.J., "Least median of squares regression," Journal of the American Statistical Association, vol. 79, no.388, pp. 871–880, 1984. DOI: 10.1080/01621459.1984.10477105.

[13] Cook R.D., "Detection of influential observation in linear regression," Technometrics, vol. 19, no.1, pp. 15–18, 1977. DOI:10.1080/00401706.2000.10485981

[14] Cook R.D., "Influential observations in linear regression," Journal of the American Statistical Association, vol. 74, no. 365, pp. 169–174, 1979. DOI: 10.1080/01621459.1979.10481634

[15] Mickey M.R., Dunn O.J., V. Clark V., " Note on the Use of Stepwise Regression in Detecting Outliers," Computers and Biomedical Research, vol. 1, no. 2 pp. 105-111, 1967. DOI:10.1016/0010-4809(67)90009-2.

[16] Gentleman J.F., Wilk M.B., " Detecting Outliers. II. Supplementing the Direct Analysis of Residuals," Biometrics, vol. 31, no. 2, pp. 387-410, 1975. DOI:10.2307/2529428.

[17] Marasinghe M.G., "A Multistage Procedure for Detecting Several Outliers in Linear Regression," Technometrics, vol.27, no. 4, pp. 941-943, 1985. DOI: 10.1080/00401706.1985.10488078.

[18] Díaz-García J.A., Gutiérrez-Jáime R., "The distribution of residuals from a general elliptical linear model," Journal of Statistical Planning and Inference , vol. 137, pp. 2347 – 2354, 2007. DOI:10.1016/J.JSPI.2006.08.003.

[19] Ellenberg J.H., "The joint distribution of the standardized least squares residual from general linear regression," Journal of the American Statistical Association, vol. 68, no. 344 ,pp. 941–943, 1973. DOI: 10.2307/2284526.

[20] Lund R.E., " Tables for an Approximate Test for Outliers in Linear Regression," Technometrics, vol. 17, no. 4, pp. 473–476, 1975. DOI: 10.1080/00401706.1975.10489374.

[21] Franklin A.G., Hariharan K .I., "Multiple linear regression," Regression Analysis: Concepts and Applications, Duxbury, 1994.

[22] Brownlee K.A., "Regression on several independent variables," Statistical Theory and Methodology en Science and Engineering, John Wiley, New York, 1965.

[23] Dodge Y., "The guinea pig of multiple regression," In H. Rieder (ed.). Lecture Notes in Statistics, No. 109, see pp. 91-118. Springer, New York,1996

[24] Daniel C., Wood F.S., "Fitting Equations to Data," John Wiley, New York, 1971.

[25] Draper N.R; Smith H., "Answers to Exercises," Applied Regression Analysis, Wiley, New York, 1981

[26] Imon A.H. M.R., "Identifying multiple influential observations in linear regression," Journal of Applied Statistics, vol. 32, no. 9 , pp. 929-946, 2005. DOI: 10.1080/02664760500163599.

[27] Alao A. N., Ayinde K., Solomn G. S.,"A Comparative Study on Sensitivity of Multivariate Tests of Normality to Outliers," ASM Science Journal, vol. 12, 2019.

[28] Arum K.C., Ugwuowo F.I., Oranye H.E ., Alakija T.O., Ugah T.E., Asogwa O.C., "Combating outliers and multicollinearity in linear regression model using robust Kibria-Lukman mixed with principal component estimator, simulation and computation,"Scientific African, vol. 9, pp. 1-17, 2023. DOI: 10.1016/j.sciaf.2023.e01566.

[29] Arum, K. C., Ugwuowo., F. I.. "Combining principal component and robust ridge estimators in linear regression model with multicollinearity and outlier," Concurrency and Computation Practice and Experience, 2022, DOI: 10.1002/cpe.6803.

[30] Lukman, A.F., Arashi, M., Prokaj, V., " Robust biased estimators for Poisson regression model: Simulation and Applications," Concurrency and Computation Practice and Experience, 2023 , DOI:10.1002/cpe.7594.

[31] Arum, K. C., Ugwuowo, F. I., Oranye, H. E., " Robust modified jackknife ridge estimator for the Poisson regression model with multicollinearity and outliers," Scientific African 17(3), 2022 , DOI: 10.1016/j.sciaf.2022.e01386.