Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 21, Issue 2(2019) ISSN$_{Print}$: 1816-9171

Identification Influential Observation for Two Stage LS with Jackknife after Bootstrap

Revised : 18\6\2019 Accepted :23\6\2019

Mohammed A. Mohammed
Al-Dewanyia Technical Institute - Al-Furat Al-awsat
Technical University/ Department of Accounting Techniques

Corresponding Author: Mohammed A. Mohammed Email: Mohammedam23@yahoo.com

Abstract|: *In the literature, there is not significant works than dealing with identifying of influential observations (IO) in the two stages least square (2SLS) models. It is a clear that when a set of high leverage points (HLPs) are exist in a data set, the traditional and the robust diagnostic approaches fails to classify correctly the IOs. This issue is because of the masking and swamping affects. In this work, we propose a new robust detection approach based on Modified Cook distance to identify the IO and to reduce the masking and swamping effects. The proposed method called modified Cook distance based on modified Generalized Studentized residuals denoted by* MGt-MCD. *In addition we suggested a new cut off point depend on the Jackknife after bootstrap. To evaluate the performance of our proposed method, a popular data set and a Monte Carlo simulation study were used. The results refer to the proposed method have good performance compared with some existing method to identify IOs in the 2SLS.*

Keywords: 2SLS, outliers, HLP, influential observations, Studentized, LMS

1 **INTRODUCTION:**coefficients simultaneous equations model (SEM) when the disturbance term across the equations are not related and the equations concerned are over identified or exactly identified. It was introduced independently by Theil, 1953 [19] and Basmann, 1957 [4]. The 2SLS estimator includes two consecutive applications of the least square estimator. The stages of 2SLS procedure are presented as follows;

First stage: Regress each of the right hand side endogenous variable in the equation to be estimated on all exogenous variables in the SEM using the least square (LS) estimator. Then, compute the predicted values for each of these endogenous variables.

Second stage: In the equation to be estimated, substitute each endogenous right hand side variable by its predicted values by using least squares estimator ([8], [15]).

The normality, homoscedasticity and no multicollinearity are the essential assumptions for LS estimators. However, the raw data are usually having heavy tail, skewed or follow a non-normal distribution ([2], [3]). This is leads to increase the standard error and then destroy the estimates. Since the presence of unusual data can violate the classical normality assumptions, robust methods is used to remedy this problem. Unusual data may come in deferent types depend on its position. We can classify unusual data into vertical outliers (lie in Y-direction) and high leverage points (lie in X-direction). Figure 1 display the types of outliers such as vertical outliers, good leverage points (GLP), bad leverage points (BLP) and influential observations ([1], [10]). And it's also shows the relationship among outliers, leverage points and influential points [16]. It is possible for a single observation may have a high effect on the predicted of a regression estimator. So, in the estimate and making the analysis, it is important to take in the account the possibility of existence of influential observations in the dataset. Belsley et al. (1980) showed that the IOs are those points that have high effects in the estimators. The high value of leverage of the observation, the more potential the point will be influential observation [5]. A point with a value close to the average on the independent variable has no impact on the line of regression line regardless of its amount on the criterion variable. On the other hand, a point that is extreme on the independent variable has the potential to affect the regression line highly. The distance of a point is based on the error of prediction for the observation. The high error for the prediction is the larger the distance. Even a point with a large distance will not have that much impact if its leverage is small. It is the combination of a point leverage and distance that identify its influence. Large number of methods has been suggested to detect an IO in linear regression model ([1], [13]). Unfortunately, most of those methods not have perfect performance due to masking and swamping effects (for more details, see [18]). Furthermore, there is not significant work for detection IO in 2SLS. In this article, we suggested a new technique to identify IOs in 2SLS. The suggested method depends on modified version of Cook distance based on modified Generalized Studentized residuals denoted by MGt-MCD. In order to assess the suggested method, a well-common data set and simulation study are performed. The correct detection ratio of the masking and the swamping were calculated to compare the study methods. Detection of influential observation with some measurements methods are discussed in Section 2. In Section 3, the Jackknife after bootstrap approach is presented. The proposed method (MGt-MCD) is given in Section 4. A numerical example and simulation study are presented in sections 5 and 6, respectively. Finally, some concluding remarks are given in Section 7.

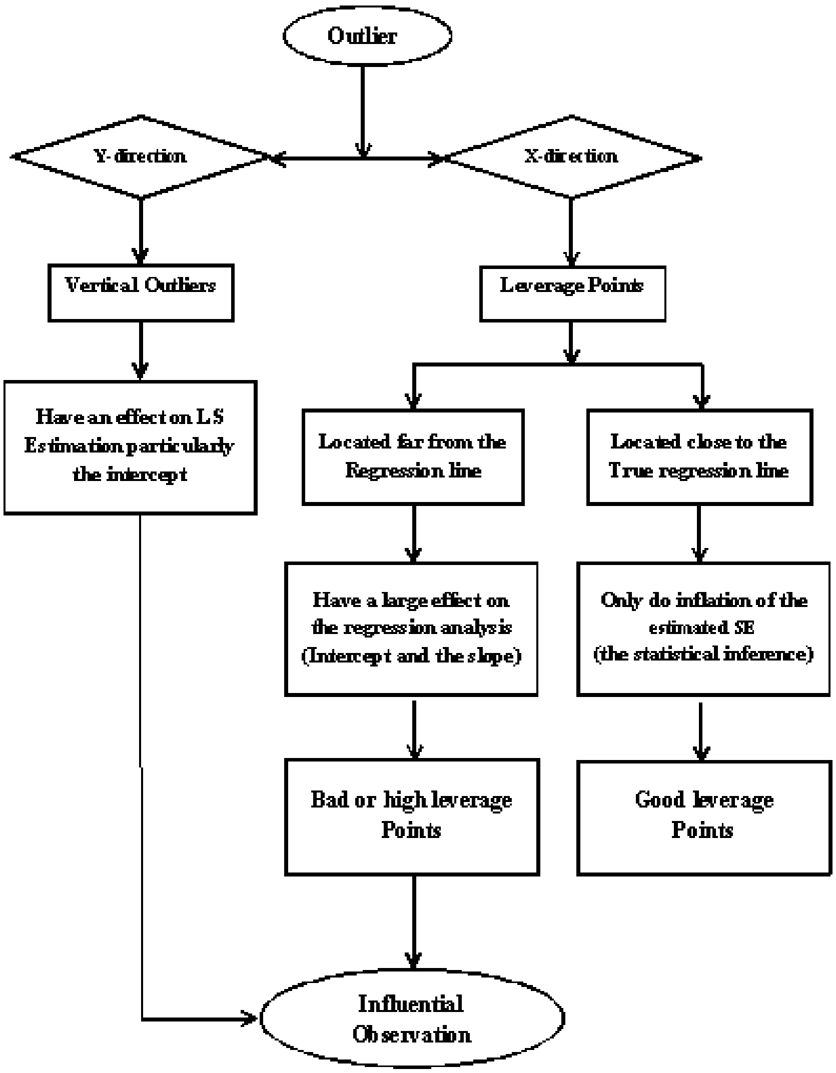

Figure 1: The Relationship among Outliers, Leverage Points and Influential Points
Source: [16]

2 DETECTION OF INFLUENTIAL OBSERVATIONS

Outliers are those outermost, or strange, data points that lie beyond the data range in either direction. In linear regression analysis, outliers can be defining as a data points that have much larger absolute residuals than all other residuals in a dataset ([2], [3]). Usually, observations with residuals lies 3σ or more away from the average of residuals is considered as outlier points. Outliers may be because of data collecting errors, data entry errors, sampling errors or any other problem with the data or “true” outliers (i.e. not an error) ([7], [11], [17]). Many approaches have been proposed to identify of outliers, however, in this study we will focus with the following identification methods.

2.1 Studentized Residuals and Modified Generalized Studentized Residual

A studentized residual sometimes referred to as (externally studentized residual) denoted as $t_i$is given as follows ([1], [13], [14], [16])

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}, i = 1,2,\dots,n \quad (1)$$

where;

$h_{ii} = x_i'(X'X)^{-1}x_i, i = 1,2,\dots,n$

$e_i = y_i - x_i\hat{\beta},$ is an i-th residuals and,

$\hat{\sigma}_{(i)}$ is a standard deviation of the residuals excluding the i-th residuals

The standardized residual follows the t-student with ($n-p,$ p is a number of covariates) degree of freedom. Rousseeuw (1990) suggested using 2.5 as a cut- off point for t_i (see [18]). The i-th deletion observation residual $\hat{\varepsilon}_{i(R)}$ is computed as,

$$\hat{\varepsilon}_{i(R)} = y_i - x_i'\hat{\beta}_{(R)}, \qquad i = 1,2,\dots,n \quad (2)$$

where, the $\hat{\beta}_{(R)}$ is the coefficient vector for the remaining groups (R), given by

$$\hat{\beta}_{(R)} = (X_R'X_R)^{-1}X_R'\,y_R \quad (3)$$

The i-th externally studentized residual t_i^* for the remaining groups "R" is given by

$$t_i^* = \frac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1-h_{ii(R)}}} \quad (4)$$

where;

$$h_{ii(R)} = x_i'(X_R'X_R)^{-1}x_i \quad i = 1,2,\dots,n \quad (5)$$

The modified generalized studentized residual (MGti) is given as follows ([1], [14]):

$$MGt_i = \begin{cases} \dfrac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1-h_{ii(R)}}}, & \text{for } i \in R \\ \dfrac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_{R}\sqrt{1+h_{ii(R)}}} & \text{for } i \notin R \end{cases} \quad (6)$$

where, $\hat{\sigma}_R$ is a standard deviation for remaining group.

2.2 MAHALANOBIS DISTANCE AND ROBUST MAHALANOBIS DISTANCE

The Mahalanobis Distance (MD) is a measurement of how far the observation x_i away from the mean the predicted variable [17]. The MD is given by;

$$MD_i = \sqrt{(x_i - \boldsymbol{x})'C_x^{-1}(x_i - \boldsymbol{x})}, i = 1,2,\dots,n \quad (7)$$

where;

$\boldsymbol{x}$ and C_x are the sample mean and sample covariance, respectively.

The cutoff-point of MD_i is equal to $\sqrt{\chi^2_{p+1,0.95}}$. Observations that exceed the cut- off point are considered as high leverage points. Rousseeuw (1985) [17] suggested a robust MD (RMD) based on minim volume ellipsoid (MVE), given as

$$RMD_i = \sqrt{(x_i - T(x))'[C(X)]^{-1}(x_i - T(x))}, i = 1,2,\dots,n \quad (8)$$

where $T(x)$ and $C(X)$ are the robust mean and covariance of the MVE, respectively.

2.3 Cook Distance Measurement

Cook Distance (CD) is a statistics measurement proposed by Cook (1977) for detecting an influential observation [9]. The CD is defined as

$$CD_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})' X' X (\hat{\beta}_{(-i)} - \hat{\beta})}{p\,\hat{\sigma}^2}, \qquad (9)$$

where, $\hat{\beta}$ is estimated parameter and $\hat{\beta}_{(-i)}$ is the estimated parameter with the *i*-th observation deleted. If $CD_i > \mathrm{F}(\alpha, p, n-p)$ the *i-th* point is considered as influential observation.
The modified version of Cook's Distance (MCD) was proposed by Cook in 1977, given by [13];

$$MCD_i = \sqrt{D_i(X'X)\frac{p(n-1)^2}{n-p}\,\hat{\sigma}^2_{(i)}} \qquad (10)$$

where, D_i is the class of norms which are location/scale. The cut- off point for MCD is defined as $2\sqrt{\frac{n-p}{n}}$.

2.4 Least Median of Squares

The least median of squares (LMS) method was suggested by Rousseeuw (1984). It is a robust approach computed by minimize the median of squared error as following ([5], [17])

$$\min Med\,(e_i^2) = \min Med\left(y_i - \sum x_{ij}\hat{\beta}_j\right)^2 \qquad (11)$$

where;
the least squared residual for the *i*-th observation (e_i) can be computed as;

$$e_i = y_i - x_i\hat{\beta} \qquad (12)$$

The LMS estimator has high break down point (BDP) of 50%, and it has little relative efficiency of 37%.

3 Jackknife after bootstrap method

The jackknife after bootstrap (JAB) approach is firstly described by Efrton (1992). The idea of the JAB approach can be summarized as following ([6], [12]):
let $\mathrm{t_i}$ is random variable with n sample of size, hence, $\mathrm{t_1, t_2, \dots, t_{i-1}, t_{i+1}, \dots, t_n}$ has a distribution which is similar to distribution of a bootstrap sample from $\mathrm{t_1, t_2, \dots, t_n}$. The JAB approach requires about e times more resamples than a usual bootstrap.
For example, for any data set, to diagnose whether an individual observation is influential or not, and to get 500 resamples without this observation, about 500 $e \approx 1500$ resamples are required. Then, we use these 500 resamples to build the sampling distribution and to diagnose the influence cut- offs. The algorithm of JAB method for detection of influential points, that proposed by Martin and Roberts, 2010, can be summarized as following [6], [12];

Step1. Suppose θ_i be the diagnostic statistic, the suitable model is fitted for original data set, and then calculate the θ_i for $i = 1,2, \dots, n$

Step2. For the original data set, create D resample with replacement.

Step3. For each observation within D resamples, obtain a subset of the samples, which do not include that observation. Therefore, there are D/e resamples attained for each observation. Compute the n values of θ_i, $i = 1,2, \dots, n$ for each of these resample, so nD/e values of θ_i are getting. Gather all nD/e values of θ into a single vector.

Step4: Determine appropriate quintiles (say 2.5% and 97.5%) of this produced bootstrapping distribution. Compare the ratios of this distribution with original θ_i, $i = 1,2, \dots, n$, values to determine the observations as influential or not.

The steps 1 to 4, should be repeated *R* times. Then, the averages and standard deviations for the number of flagged observations for all these *R* generated can be computed. It has to consider, the JAB algorithm runs just one time with real dataset. If the original dataset include influential data points, these data points will potentially show several times in the generated sampling distributions. This is lead to; quantities computed from these samples will not be acceptable for comparison. In respect of compute the suitable quantities, the cut-offs will determine from the sampling distribution using resamples not include the point concerned.

4 The proposed method (MGt-MCD)

The classical techniques denoted by STAND-MD and STUDENT-MD draw the standardized residuals and studentized residuals against the Mahalanobis distance, respectively. Rousseeuw and Van Zomeren (1990) suggested a robust detection method to classify dataset into clean data, vertical outliers, GLB and bad leverage points [18]. The Rousseeuw and Van Zomeren's method based on the LMS versus RMD denoted by (LMS-RMD). The classical methods are not sufficient technique to identify IOs due to it based on the non-robust method. In addition, the LMS-RMD is not very effective technique to classify the data points into respective categories since it is based on the RMD, which suffers from swamping effects. To rectify the darkness of existing method we proposed a new technique by combining the MGti against MCD denoted by MGt-MCD. The cut- off points of MGt-MCD is based on the Jackknife after bootstrap ([6], [12]) that presented in Section 4.

The rules for classification observation using the MGt-MCD method are displayed as following.

- **i.** The observation "i" *is* declared as clean data if $|MGt_i| \leq 2.5$ and $MCDi \leq$ JAB cut $-$ off.
- **ii.** The observation "i" *is* declared as vertical outlier if $|MGt_i| > 2.5$ and $MCDi \leq$ JAB cut $-$ off.
- **iii.** The observation "i" *is* declared as a GLP if $|MGt_i| \leq 2.5$ and $MCDi >$ JAB cut $-$ off.
- **iv.** The observation "i" *is* declared as a BLP if $|MGt_i| > 2.5$ and $MCDi >$ JAB cut $-$ off.

were, "JAB cut $-$ off " is a JAB cut- off point of MCD that is computed using R language.

5 Hawkins Brado Kass Dataset

The Hawkins, Bradu and Kass (HBK) dataset is an artificial data set introduced by Hawkins et. al. in 1984 [17]. The dataset has 75 observations and contents one response variable with 3 independent variables. This data set includes 14 high HLPs (cases 1-14), however, (cases1-10) are BLPs and (cases11-14) are GLPs. In this example we are applying the STAND-MD, STUDENT -MD, LMS-RMD and the proposed method MGt-MCD to diagnostic influential observations in both stages of the 2SLS model for HBK dataset. We suppose that the first stage has two independent variables, x_1 and x_2 and the second stage has $\hat{x}_1$ and x_3. Figures 2 and 3 present plot for diagnostic methods for the first and second stages, respectively. And the tables 1 and 2 present the values of the diagnostic method for the first 14th observations. The result of the example pointed that the MGt-MCD has the best performance followed by LMS-RMD for identifying the IOs. It interesting to see; only the MGt-MCD identify correctly the good and bad leverage points in both stages as shown in figures 2 and 3. Tables' 1and 2 display the values of the diagnostic methods for the first and second stages of 2SLS for the HBK dataset. Numbers in **bold style** represent the values those exceed its cut-off point which are identified as IOs. Where, the cut-off points for the methods put in the parentheses.

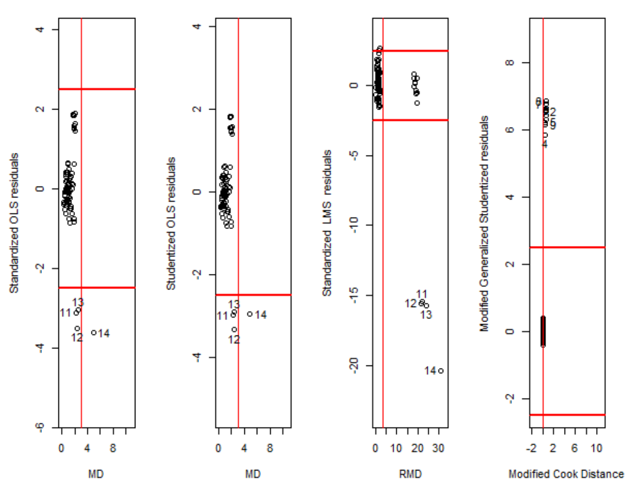

Figure 2: plot for diagnostic method for the first stage for HBK dataset

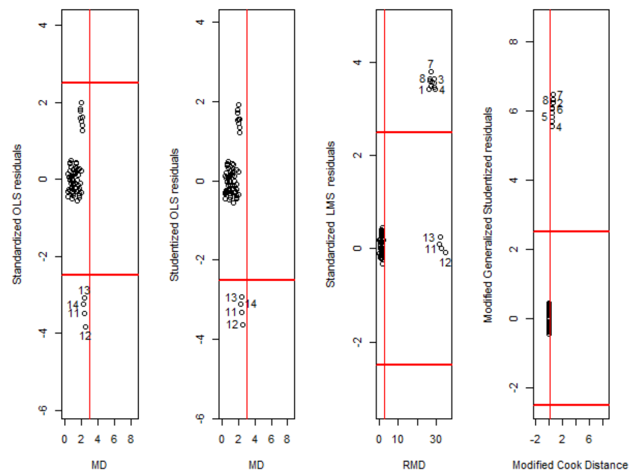

Figure 3: plot for diagnostic method for the second stage for HBK dataset

Table 1: The values for diagnostic method for the first stage for HBK dataset

Number of observation	MD (3.057)	RMD (3.618)	MCD (0.055)	Studentized (2.5)	LMS (2.5)	MGti (2.5)
1	1.91	**18.14**	**0.51**	1.49	-0.02	**6.43**
2	1.81	**18.69**	**0.53**	1.82	0.26	**6.56**
3	2.12	**18.87**	**0.54**	1.57	0.20	**6.59**
4	1.93	**19.68**	**0.43**	1.45	-1.20	**5.87**
5	1.95	**19.48**	**0.49**	1.55	-0.51	**6.24**
6	2.14	**19.08**	**0.50**	1.42	-0.34	**6.33**
7	2.00	**19.38**	**0.57**	1.83	0.55	**6.78**
8	1.84	**18.07**	**0.58**	1.80	0.83	**6.87**
9	1.84	**18.93**	**0.47**	1.55	-0.57	**6.15**
10	1.72	**17.94**	**0.54**	1.80	0.51	**6.66**
11	2.23	**22.18**	0.00	**-2.98**	**-15.45**	-0.10
12	2.46	**21.71**	0.00	**-3.33**	**-15.57**	-0.15
13	2.48	**24.23**	0.00	**-2.90**	**-15.75**	0.04
14	**4.90**	**30.95**	0.00	**-2.95**	**-20.38**	-0.04

Table 2: The values for diagnostic method for the second stage for HBK dataset

Number of observation	MD (3.058)	RMD (3.159)	MCD (0.056)	Studentized (2.5)	LMS (2.5)	MGti (2.5)
1	1.89	**26.27**	**0.50**	1.55	**3.42**	**6.08**
2	1.86	**26.72**	**0.53**	1.73	**3.61**	**6.24**
3	2.06	**28.89**	**0.49**	1.56	**3.64**	**5.96**
4	2.18	**29.41**	**0.42**	1.23	**3.42**	**5.56**
5	2.03	**28.91**	**0.47**	1.46	**3.56**	**5.84**
6	2.13	**27.24**	**0.50**	1.55	**3.50**	**6.06**
7	2.01	**27.09**	**0.57**	1.93	**3.80**	**6.48**
8	1.85	**26.70**	**0.54**	1.79	**3.65**	**6.33**
9	2.12	**28.72**	**0.44**	1.34	**3.46**	**5.70**
10	2.13	**28.00**	**0.48**	1.55	**3.58**	**5.96**
11	2.42	**32.67**	0.02	**-3.31**	0.01	0.32
12	2.54	**34.70**	0.02	**-3.62**	-0.09	0.30

13	2.41	**31.91**	0.03	**-2.93**	0.24	0.41
14	2.29	**31.73**	0.02	**-3.11**	0.10	0.35

6 Monte Carlo Simulation Study

In this section, a simulation studies are organized to evaluate the performance of the proposed method for detection of IOs in the two stages LS model. The methods of the study are compared based on the average number correctly detected of IOs. A good diagnostic method is the one that have detected closer number of IOs that created in the data set with lowest number of masking and swamping ratio. The experimental design is described as follows;

We generate randomly 5 explanatory variables following the normal distribution as $x_j \sim N(0, 1),\ j = 1,2, \dots ,5$. The response variables are generated according the following equations;

First stage; $$y_1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u_1 \quad (13)$$

Second stage; $$y_2 = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_4 + \tilde{\beta}_3 x_5 + u_2 \quad (14)$$

where, β_1, β_2 and β_3 are assumed equal to one, and $\tilde{\beta}_1, \tilde{\beta}_2$ and $\tilde{\beta}_3$ are assumed equal to 0.5. The error term u_1 and u_1 are distributed as $N(0, \sigma_i^2)$. Two sizes of samples are setting equals to ($n = 40$ and 100). The generating data are contaminated with different percentages of outliers [$\alpha = 0.05, 0.10,$ and 0.20] in both X and Y directions. To create outliers in X and Y direction, the first α % percent of clean data were replaced by fixed value equal to 10. For consistency of results, the Monte Carlo simulation studies are repeated 500 times. The STAND-MD, STUDENT-MD, LMS-RMD and the MGt-MCD methods are applying to detect the outliers and IOs in the generated dataset. Tables 3 to 6 present the results of simulation study. The results indicate that our proposed method (MGt-MCD) has the supreme performances with excellent average of detection for outliers and influential observations. Moreover, the MGt-MCD diagnostic method has the lowest ratio of masking and swamping effects. The LMS-RMD has moderate performance of detection of IOs but it has the largest ratio of masking and swamping effects. It clear to see the STAND-MD, STUDENT-MD methods failed to detect influential observations and have not any swamping ratio.

Table 3: The average number of identification of outliers for diagnostic methods when $n = 40$, $u_1 \sim N(5, 10)$ and $u_2 \sim N(5, 10)$

Contaminated percentage	Method	*First stage*			*Second stage*		
		Correct identify	*Masking*	*Swamping*	*Correct identify*	*Masking*	*Swamping*
5 % 2 outliers	STAND-MD	50%	50%	0%	0%	0%	0%
	STUDENT-MD	50%	50%	0%	0%	0%	0%
	LMS-RMD	100%	0%	0%	100%	0%	0%
	MGt-MCD	100%	0%	0%	100%	0%	0%
10 % 4 outliers	STAND-MD	50%	50%	0%	0%	0%	0%
	STUDENT-MD	50%	50%	0%	0%	0%	0%
	LMS-RMD	100%	0%	0%	75%	25%	25%
	MGt-MCD	100%	0%	0%	100	0%	0%
% 8 outl	STAND-MD	50%	50%	0%	100%	0%	0%
	STUDENT-MD	50%	50%	0%	100%	0%	0%

	LMS-RMD	50%	50%	50%	25%	75%	50%
	MGt-MCD	100%	0%	0%	100	0%	0%

Table 4: The average number of identification of outliers for diagnostic methods when $n = 100$, $u_1 \sim N(5, 10)$ and $u_2 \sim N(5, 10)$

Contaminated percentage	Method	*First stage*			*Second stage*		
		Correct identify	*Masking*	*Swamping*	*Correct identify*	*Masking*	*Swamping*
5 % 5 outliers	STAND-MD	20%	80%	0%	0%	0%	0%
	STUDENT-MD	20%	80%	0%	0%	0%	0%
	LMS-RMD	80%	20%	20%	80%	20%	20%
	MGt-MCD	100%	0%	0%	100%	0%	0%
10 % 10 outliers	STAND-MD	0%	100%	0%	0%	100%	0%
	STUDENT-MD	0%	100%	0%	0%	100%	0%
	LMS-RMD	60%	40%	25%	50%	50%	50%
	MGt-MCD	100%	0%	0%	100	0%	0%
20 % 20 outliers	STAND-MD	0%	100%	0%	100%	0%	0%
	STUDENT-MD	0%	100%	0%	100%	0%	0%
	LMS-RMD	25%	75%	50%	25%	75%	50%
	MGt-MCD	100%	0%	0%	100	0%	0%

Table 5: The average number of identification of influential observations for diagnostic methods when $n = 40$, $u_1 \sim N(5, 10), u_2 \sim N(5, 10)$ and *X's have HLPs*

Contaminated percentage	Method	*First stage*			*Second stage*		
		Correct identify	*Masking*	*Swamping*	*Correct identify*	*Masking*	*Swamping*
5 % 2 outliers	STAND-MD	0%	100%	0%	0%	100%	0%
	STUDENT-MD	0%	100%	0%	0%	100%	0%
	LMS-RMD	0%	100%	50%	0%	100%	50%
	MGt-MCD	100%	0%	0%	100%	0%	0%
10 % 4 outliers	STAND-MD	0%	100%	0%	0%	100%	0%
	STUDENT-MD	0%	100%	0%	0%	100%	0%
	LMS-RMD	20%	80%	50%	25%	75%	50%
	MGt-MCD	100%	0%	0%	100	0%	0%
20 % 8 outliers	STAND-MD	0%	100%	0%	0%	100%	0%
	STUDENT-MD	0%	100%	0%	0%	100%	0%
	LMS-RMD	50%	50%	50%	25%	75%	50%
	MGt-MCD	100%	0%	0%	100	0%	0%

Table 6: The average number of identification of outliers for diagnostic methods when $n = 100$, $u_1 \sim N(5, 10)$ and $u_2 \sim N(5, 10)$

Contaminated percentage	Method	*First stage*			*Second stage*		
		Correct identify	*Masking*	*Swamping*	*Correct identify*	*Masking*	*Swamping*
out	STAND-MD	0%	100%	0%	0%	100%	0%

	STUDENT-MD	0%	100%	0%	0%	100%	0%
	LMS-RMD	0%	100%	20%	20%	80%	20%
	MGt-MCD	100%	0%	0%	100%	0%	0%
10 % 10 outliers	STAND-MD	0%	0%	0%	0%	0%	0%
	STUDENT-MD	0%	0%	0%	0%	0%	0%
	LMS-RMD	50%	50%	25%	60%	40%	50%
	MGt-MCD	100%	0%	0%	100	0%	0%
20 % 20 outliers	STAND-MD	0%	100%	0%	100%	0%	0%
	STUDENT-MD	0%	100%	0%	100%	0%	0%
	LMS-RMD	50%	50%	25%	75%	25%	50%
	MGt-MCD	100%	0%	0%	100	0%	0%

7 CONCLUSIONS

The results of numerical data and simulation study indicate the following:

The classical methods such as STAND-MD and STUDENT-MD failed absolutely to diagnose the influential observations. However, the classical methods did not suffering from swamping effects. Whereas, the LTS-RMD method has moderate performance for detecting outliers and influential

observations but it is suffering from masking and swamping effects. It is clear to see that the proposed method (MGt-MCD) is successfully detects the influential observations with reducing masking and swamping effects. It is easy to conclude that the MGt-MCD method has a supreme performance for detecting influential observations among all of the existing diagnostic methods without any ratios of masking and swamping effects.

8 REFERENCES

[1] Alguraibawi, M., Midi, H. and Imon, A.H.M., 2015. A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering*, *2015*.

[2] Atkinson, A. and Riani, M., 2012.*Robust diagnostic regression analysis*. Springer Science & Business Media.

[3] Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association* 89: pp.1329–1339

[4] Basmann, R. 1957. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25, pp.77–83.

[5] Belsley, D. A., Kuh, E., and Welsch, R. E., (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.

[6] Beyaztaş, U., 2012. *Jackknife-after-bootstrap method for detection of outliers and influential observations in linear regression models* (Doctoral dissertation, DEÜ Fen Bilimleri Enstitüsü).

[7] Bondell, H.D. and Stefanski, L.A., 2013. Efficient robust regression via two-stage generalized empirical likelihood. *Journal of the American Statistical Association*, *108*(502), pp.644-655.

[8] Chatterjee, S. and Hadi, A.S., 2015. *Regression analysis by example*. John Wiley & Sons.

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

[10] Imon, A.H.M.R., 2002. Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies*, *3*, pp.207-218.

[11] Klym, A.H., 2007. *Detecting Outliers and Influential Data Points in Receiver Operating Characteristic (ROC) Analysis*(Doctoral dissertation, University of Pittsburgh).

[12] Marubini, E. and Orenti, A., 2014. Detecting outliers and/or leverage points: a robust two-stage procedure with bootstrap cut-off points. *Epidemiology, Biostatistics and Public Health*, *11*(3).

[13] Habshah, M., Norazan, M.R. and Rahmatullah Imon, A.H.M., 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, *36*(5), pp.507-520.

[14] Midi, H.A.B.S.H.A.H. and Mohammed, A.M., 2015. The Identification of Good and Bad High Leverage Points in Multiple Linear Regression Model. *Mathematical Methods and System in Science and Engineering*.

[15] Mishra, S.K., 2008. Robust Two–Stage Least Squares: Some Monte Carlo Experiments. *Journal of Applied Economic Sciences (JAES)*, *3*(06), pp.434-443.

[16] Mohammed, E.T.H., 2012. Robust estimation in simultaneous equations model. *CU Theses*.

[17] Rousseeuw, P.J. and Leroy, A.M., 1987. *Robust regression and outlier detection* (Vol. 1). New York: Wiley.

[18] Rousseeuw, P.J. and Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, *85*(411), pp.633-639.

[19] Theil, H. 1953a. Repeated least-squares applied to a complete equation systems. Mimeo. The Hague: Central Planning Bureau.