Improvise Group Diagnostic Potential Measure for Multivariate Normal Data

Hassan S. Uraibi[(1)] **Sawsan A. Alhussieny**[(2)]

hassan.uraibi@qu.edu.iq Stat.post22@qu.edu.iq

Dept. of Statistics, College of Administration and Economics

University of Al-Qadisiya, IRAQ

Abstract

The diagnosing of outliers is considered a very significant topic in many scientific fields. The existence of outliers in the dataset leads to the breakdown of the method estimator. There are numerous types of outliers that classified according to the nature of the data, as the statistical literature showed. Consequently, the researchers focused on identifying the type of outliers of statistical models by utilizing two diagnostic procedures, individual and group. The individual procedure, unfortunately, neglects the impact of the phenomenon that is masking and swamping, while the second procedure was unable to eliminate this phenomenon completely, but rather decrease the rates of its appearance. The present paper is suggesting the development of one of the famous group diagnostic methods that are so-called (IDRGP) through making use of an RMVN location and scale matrix instead of MVE to decrease the impact of (swamping). The performance of the proposed method that is denoted as (IDRGP.RMVN) has been tested with a certain number of simulation studies and applied with real data. The outcomes show that the performance of our suggested method is more efficient than (IDRGP.MVE) to decrease the swamping points where the sample size is large in the presence of all kinds of outliers.

Keywords: masking and swamping, outliers, IDRGP.MVE, IDRGP.RMVN

1. Introduction

The assumption of normal distribution is formed the backbone of the classical school of statistics and the violation this assumption leads to different interpretations. Peirce (1852) observed that the source of some observations is abnormal and it is quite different from the rest of the observations that are supposed to be normally distributed (Hampel, et al., 1986). Therefore, it is perhaps take different shape that it might be skewed to the left or to the right or with a heavy tail. Geary (1947) made a controversial around the assumption of the normal distribution again when he mentioned that the idea of the normality distributed of observations is just a myth and there is no normal distribution and it will not be (Huber, 1981)

The researchers found that the presence of outliers is the one of important reasons that some observations deviate from normal distribution assumption. So, it is important to diagnostic these abnormal observations being considered far away from the center of the gathering bulk of data. Rousseuw and Zomeren (1990) defined outliers as observations that lie away from most of the rest of the data as well as it constitutes (1%) to (10%) out of any group of data in the real world. Lately, some researchers displayed that this ratio may be raised to more than (25%) and less than (50%), but it is inevitable even if this data is of high quality (Hample, 1986). Huper (1981) pointed out that the presence of at least one of the outliers in the data group leads to the breakdown of the statistical estimator.

Thus, it is obvious that the diagnostic of the outliers in the data set is quite important. The outliers are labelled in relation to the statistical model into different types. Some of them occur in the random errors are called outliers, while those that appeared in independent variables for the multiple linear regression model are called Leverage Points (LP), this conception also includes multivariate analysis such as multi-response regression, factor analysis, discriminatory analysis, etc.

Belsley et al. (1980) identified another type of observations, calling them influential observations (IO) which are influential on statistical samples that depend on the influence of independent variables on one or more of the variables used. Hadi (1992) suggested deletion measure to detect high leverage points. It is well-known that this measure is the so-called Hadi's potential measure that was considered as one of the single diagnostic methods in the statistical literature.

Unfortunately, the previous methods conceal in their folds the wrong diagnosis when its methods detect one or more than one observation as outliers but it's not, this phenomenon is called (swamping). On the other hand, may these methods suffering from the masking phenomenon in which the detected outliers probably overshadow other outliers, therefore the certain diagnostic method could not detect the outliers that masked by other outliers.

The Generalize Potential (GP) measure introduced by Imon (2002) as a group deletion measure to get rid of the effect of masking and swamping. Unlucky, this procedure is not sufficient and therefore Midi et al. (2009) mentioned that GP could not identify the exact number of leverage points as well as still suffering from the impact of masking and swamping. Consequently, they proposed utilized from Minimum Volume Ellipsoid (MVE) (Rousseeuw, 1984) to build a new algorithm which is a so-called Diagnostic Robust Generalized Potential measure (IDRGP). We observed that IDRGP.MVE may treat the problem of identifying the exact number of leverage points, but it is not adequately effective in reducing the number of masking and swamping or get rid of its effects.

The MVE algorithm is not feasible option particularly with high dimensional data, because it is a time-consuming even with the Fast algorithm of it that suggested by

Rousseeuw and Van Driessen(1999), see (Khan et al.,2007a; Khan et al. 2007b; Uraibi and Midi,2019). Olive and Hawkins (2010) introduced Reweighted MultiVraite Normal (RMVN) as a robust, Fast, and Consistent concentration algorithm to produce a robust location and scale estimator. Due to the aspects of RMVN, we thought that it is more relevant to IDRGP than MVE. It is well known that IDRGP.MVE algorithm relies on Robust Mahanalobis Distance (RMD) that integrated with MVE estimators. In this paper, a slight development to the IDRGP is proposed and we call it IDRGP.RMVN by incorporating RMVN with RMD instead of MVE.

This paper is set to present the Hat Matrix in Section 2. Section 3 explains the IDRGP(MVE) measure, Section 4 presents IDRGP(RMVN) measure, Section 5 illustrate simulation study, section 6 the numerical example to assess the performance of the IDRGP(RMVN) method, finally the conclusion is viewed in section 7.

2. Projection or Hat Matrix

It is also known as weight matrix and it is also known as (Hat matrix) and it is used in determining the observations rows x containing the outliers and known as the regression analysis as a detecting scale for the existence (LP) and the final mathematical form is :-

$$w = x\left(x'x\right)^{-1}x' \tag{1}$$

And the diagonal for the matrix is w and it can be written as follows:

$$w_{\mathrm{ii}} = x_i\left(x'x\right)^{-1}x' \tag{2}$$

These elements has useful features especially its values ranges between (1,0) and the summation of w_{ii} equals p. In addition to that we can prove that w_{ii} which contain LP to the case i^{th} which is the measurement of the distance between x values to the i^{th} case and the mean of x values for all cases n, thus the big value to w_{ii} refers to w_{ii} that the case i^{th} is far away of the centre of the observations of the *x* variable both, and the mean of this matrix is written as follows:

$$\bar{w} = \frac{\sum_{-1}^{n} w_{ii}}{n} = \frac{p}{n} \tag{3}$$

Where p is the number of the variables and n is the whole number for the observations.

(Hoagline and Wehhsch) suggested at 1978 that the cutter in the presence of LP in a variable without one else and the value of $w_{\text{ii}} > \frac{2p}{n}$ also there is a basis which is three times double more than the mean that has been presented by Velleman and Wehhsch at 1981 to diagnose LP when $w_{\text{ii}} > \frac{3p}{n}$.

3. IDRGP(MVE) Measure

This method presented by Mohammed et al. (2015) to improve the performance of DRGP(MVE). they noticed that there is an impact of the swamping and masking cases when the percentage of HLP is between 5% and 10%. They pointed out that the diagnostic of HLP in the second step of DRGP(MVE) algorithm in which the partial matrix called D has not been checked correctly. Since, the algorithm of DRGP(MVE) can be summarized as follows,

1- Computing the location $\boldsymbol{T_R(x)}$ and scale $\boldsymbol{C_R(x)}$ parameter estimator of MVE

2-Measuring robust Mahalanobies distance of full dataset according to the following equation:

$$\boldsymbol{RMD_i(MVE) = \sqrt{(x - T_R(x))' C_R(x)^{-1} (x - T_R(x))} \quad i = 1, 2, \dots, n} \quad (4)$$

The suspected observations as LP should be putting n the D matrix, while the rest of the observations put in the R matrix and the obtain the GP (Imon,2002) algorithm outcomes P_{ii} by using the following equation,

$$P_{\text{ii}} = \begin{cases} W_{\text{ii}}^{(-\text{D})} & \forall \text{i} \in \text{D} \\ \frac{W_{\text{ii}}^{(-\text{D})}}{1 - W_{\text{ii}}^{(-\text{D})}} & \forall \text{i} \in \text{R} \end{cases} \quad (5)$$

any of the values of P_{ii}that exceeds the cutoff point,

$$median(P_{\text{ii}}) + 3MAD(P_{\text{ii}}) \quad (6)$$

is considered an outlier.

Thus, they suggested adding a further step to the algorithm through the diagnostic of LP by the use of the hat matrix and then compared with the first diagnosis. Within, they are going to compare what is diagnosed as HLP being a final result for the first algorithm that he suggested as partial matrix D_2 then comparing with what is founded in the second step that resulted in the partial matrix D which means $D_2$and D and as follows,

1- If the observations diagnosed as HLP are the same as D_2 and D thus the algorithm will the announcement of this diagnosis and then stop.
2- If the number of HLP in D_2 are more than those in D, then the algorithm works on move those observations that are not matched with D to the matrix R one by one according to P_{ii} value. if The value of P_{ii} for certain observation exceeds the cutoff point in equation (6) stay in D_2 matrix, otherwise it move to R matrix.
3- If the number of the HLP in $D_2$is less than the number of what is diagnosed in D which means new observations that have not been diagnosed before, that the algorithm works on merging between D and D_2. Re-checking the suspected observations by using the generalized potential measure P_{ii} is crucial to confirm whether these observations are HLP or clean. So, clean observation should be moved to R matrix.

4. The IDRGP(RMVN) Measure

The contribution of the suggested method is to incorporate the Reweighted Multivariate Normal estimators (RMVN) estimators instead of (MVE) estimators within the DRGP algorithm. Olive and Hawkins (2010) proposed the RMVN method to reweight multivariate normal estimators by using a fast and consistent algorithm which is having a high breakdown point. In the first two stages, the estimators of two location and scale have been computed, the DGK (Devlin et al., 1981) and Median Ball (MD) (Olive,2004). The DGK and MB are fast concentration algorithms that could be convergence within 5 to 10-steps.

1- The algorithm starts with the classical mean and variance as initial two estimators that are denoted as $(T_{0,1}, C_{0,1})$ respectively, and then five steps of concentration algorithm DGK is sufficient to converge and give robust estimators. In each iteration, Mahala Nobis Distant, new location and scale matrix values are computed. The estimators of each iteration are calculated from the data that poses its MD is not more than the median of MD values at that step.
2- The concentrated algorithm of MB begins with median and identity matrix as location and scale estimators, and the follow the similar steps of DGK except in each step the median is computed instead of mean.
3- Suppose that $(T_{5,DGK}, C_{5,DGK})$ and $(T_{5,MB}, C_{5,MB})$ are the final estimator of DGK and MB respectively. the FCH location and scale estimators can be obtained by

$$T_{FCH} = \begin{Bmatrix} T_{5,DGK} & if \sqrt{|C_{5,DGK}|} < \sqrt{|C_{5,MB}|} \\ T_{5,MB} & Otherwise \end{Bmatrix} \quad (7)$$

$$\mathrm{C_{FCH}} = \begin{Bmatrix} \frac{MED\left(MD_i^2((T_{5,DGK}, C_{5,DGK}))\right)}{\chi^2_{(p,0.5)}} \times C_{5,DGK}, & if \sqrt{|C_{5,DGK}|} < \sqrt{|C_{5,MB}|} \\ \frac{MED\left(MD_i^2((T_{5,MB}, C_{5,MB}))\right)}{\chi^2_{(p,0.5)}} \times C_{5,MB}, & \text{Otherwise} \end{Bmatrix} \quad (8)$$

where $|\blacksquare|$ stand for the determinant of scale matrix and MD is the traditional Mahalanobis Distance.

Let $(\hat{T}_1, \hat{C}_1)$ be the traditional estimator applied to n_1 cases with $MD_i^2[(\ \mathrm{T_{FCH}}, \mathrm{C_{FCH}})\ ] \leq \chi^2_{(p,0.975)}$,and let $q_1 = min\left\{\frac{(0.5\times 0.975\times n)}{n_1}, 0.995\right\}$

So, the first standard reweighting of MVN data is,

$$C^{(1)}_{RMVN} = \frac{MED\left(D_i^2(T_{FCH}, C_{FCH})\right)}{\chi^2_{(p,q_1)}} \times \mathrm{C_{FCH}} \quad (9)$$

The new estimators $(\ \mathrm{T_{FCH}}, C^{(1)}_{RMVN})$ are applied to n_2 cases with $MD_i^2\left[(\ \mathrm{T_{FCH}}, C^{(1)}_{RMVN})\ \right] \leq \chi^2_{(p,0.975)}$, and let$q_2 = min\left\{\frac{(0.5\times 0.975\times n)}{n_2}, 0.995\right\}$, the RMVN estimator can be found as follows,

$$C^{(2)}_{RMVN} = \frac{MED\left(D_i^2\left(T_{RMVN}, C^{(1)}_{RMVN}\right)\right)}{\chi^2_{(p,q_2)}} \times C^{(1)}_{RMVN} \quad (10)$$

The algorithm of DRGP (RMVN) measure can summarize as follows,

1. Computing the locationT_{RMVN} and scale $C^{(2)}_{RMVN}$ estimators.
2. Calculating Mahalanobis Distance MD by Eq. (10) and the $i^{th} MD_i(RMVN) > \sqrt{\chi^2_{(p,0.95)}}$ then the i^{th} row is having the suspected observations as HLP.

$$MD_i(RMVN) = \sqrt{(x - T_{RMVN}(x))' {C^{(2)}_{RMVN}}^{-1'} (x - T_{RMVN}(x))} \quad (10)$$

3. Deletion D rows from matrix X of original data, where $D = \left\{MD_i(RMVN) > \sqrt{\chi^2_{(p,0.95)}}\right\}$ is the rows index and then put the deletion rows in X_D matrix, while the remining rows will be in X_R matrix, and then the algorithm of GP should produce P_{ii}.
4. The remaining steps are similar to the algorithm of IDRGP (.)

5.Simulation Study

Let's suppose the multiple linear regression be as follows:

$$y = x\beta + e$$

Where x is $n \times p$ design matrix that generates from multivariate normal distribution with mean equals to zero and standard deviation equivalents to $\sigma = \rho^{|i-j|}$ which means, $x \sim N(0, \rho^{|i-j|})$, where $p = 15$, n is the generated sample that will take different number of observations, $n = \{50,70,100,200,300\}$, β is the identity vector of this model

$$\beta = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{15x1} \qquad (12)$$

and e is random error term which is distributed normal with zero mean and 2 standard deviation.

In order to make sure of the diagnosis efficient of comparative methods we contaminate the simulated data with different proportions of outliers, $\alpha = (0.05, 0.10, 0.15)$ as follows:

1-Contaminating the design matrix of each sample by α BLP in the presence of one HLP. That is by multiplying the first three rows of the second variable to the fifth variable by the number 10, and multiplying the maximum value of the first variable by the number 10, as well as what corresponds to it in the response variable Y.

2-Contaminating the random errors of each sample by α Vertical Outliers (V.O) in the presence of one HLP. The V.O, sare generated from a chi-square distribution with (10) degree freedom.

3. Contaminating both design matrix and random errors α LP & Vertical Outliers (V.O) in the presence of one HLP.

The main reason of including single HLP is all cases of simulation study is to consider the Phenomena of masking and swamping.

Let $\lambda_i$is random variable, where$i = 1,2, \dots, n$, and let the $O = \{\lambda_1, \dots, \lambda_\delta\}$ are the outlying observations, such that $(\delta = \alpha \times n)$and α are the number and the percentage of outlying observations, respectively. The clean observations would be $C = \{\lambda_{\delta+1}, \dots, \lambda_n\}$.

Suppose that E_j is the outlying cases that detected by certain diagnostic method, where $1 \le j \le \delta^*$, δ^* either $(\delta + b)$ or $(+b)$, such that h and b are integer number, $[0 \le b < n]$and $[0 \le h < \delta]$,

Consequently, $\lambda_b \in C$and $\lambda_h \in O$and we can conclude that the exact detection will be happen when $(\delta^* = \delta)$ in which no swamping cases $(b = 0)$ nor masking cases $(h = 0)$. However, the certain method would be having swapping cases where $(\delta^* > \delta)$ and masking where$(\delta < \delta - h)$.

The performance of our proposed are compared with others over all (1000) datasets for each simulation case. The best diagnostic method is the one that has average of correct diagnostic closer to δ (correct), lower average of b (swap) and reduced the computation time.

Table -1-averages of the **correct** diagnosis, S**wap** and the Time of computation, respectively, forthree cases of simulation when $\alpha = 0.05$ and different sample sizes.

contamination	**n**	**IDRGP(MVE)**			**IDRGP(RMVN)**		
		correct	**Swap**	**Time**	**correct**	**Swap**	**Time**
LP	50	**3.904**	**3.367**	**0.491**	**3.912**	**1.710**	**0.046**
	100	**5.912**	**2.741**	**0.748**	**5.912**	**1.189**	**0.091**
	200	**10.902**	**3.495**	**1.328**	**10.905**	**1.704**	**0.192**
	300	**15.900**	**4.562**	**1.956**	**15.913**	**2.319**	**0.310**
	500	**25.7**	**6.400**	**3.848**	**25.7**	**3.100**	**0.617**
	1000	**50.9**	**13.800**	**8.035**	**50.9**	**7.700**	**1.701**
V.O.	50	**3.859**	**3.403**	**0.503**	**3.909**	**1.639**	**0.047**
	100	**5.942**	**4.088**	**0.764**	**5.938**	**1.180**	**0.094**
	200	**10.947**	**7.680**	**1.346**	**10.936**	**1.680**	**0.198**
	300	**15.946**	**11.781**	**1.959**	**15.925**	**2.407**	**0.316**
	500	**25.85**	**15.850**	**3.422**	**25.75**	**3.150**	**0.614**
	1000	**50.85**	**36.750**	**8.019**	**50.85**	**6.950**	**1.714**
LP &V.O.	50	**3.942**	**2.429**	**0.494**	**3.942**	**1.813**	**0.045**
	100	**5.956**	**2.17**	**0.753**	**5.956**	**1.194**	**0.091**
	200	**10.958**	**4.196**	**1.336**	**10.958**	**1.675**	**0.193**
	300	**15.957**	**6.487**	**1.955**	**15.958**	**2.236**	**0.311**
	500	**26**	**11.250**	**3.355**	**26**	**4.050**	**0.587**
	1000	**50.9**	**21.700**	**7.865**	**50.9**	**7.400**	**1.654**

Table(1) shows that when the pollution average is (0.05) for all the outliers (LP, V.OUT, LP &V.OUT) , we noticed that our suggested way is more preferable in its accuracy and the reduction of the incorrect diagnosis in a measuring time faster than

IDRGP(MVE) when the sizes of the samples are (50, 100, 200, 300) yet it losses the aspect of reducing the average of diagnosing the incorrect swap at big sizes of samples.

Table -2-averages of the **correct** diagnosis, **Swap** and the **Time** of computation, respectively, for three cases of simulation when $\alpha = 0.10$ and different sample sizes.

contamination	**n**	**IDRGP(MVE)**			**IDRGP(RMVN)**		
		correct	**Swap**	**Time**	**correct**	**Swap**	**time**
LP	50	**5.821**	**2.853**	**0.488**	**5.827**	**1.226**	**0.045**
	100	**10.844**	**1.743**	**0.742**	**10.850**	**0.718**	**0.090**
	200	**20.779**	**1.926**	**1.308**	**20.788**	**0.903**	**0.186**
	300	**30.726**	**2.399**	**1.955**	**30.756**	**1.235**	**0.304**
	500	**50.9**	**3.300**	**3.640**	**50.9**	**1.800**	**0.661**
	1000	**99.9**	**3.800**	**7.862**	**100.1**	**2.700**	**1.611**
V.OUT	50	**5.624**	**3.098**	**0.499**	**5.743**	**1.341**	**0.046**
	100	**10.872**	**3.294**	**0.755**	**10.877**	**0.748**	**0.092**
	200	**20.862**	**7.296**	**1.354**	**20.828**	**0.852**	**0.198**
	300	**30.882**	**11.201**	**1.962**	**30.817**	**1.322**	**0.317**
	500	**50.850**	**16.850**	**3.474**	**50.850**	**1.850**	**0.629**
	1000	**100.7**	**42.300**	**8.282**	**100.4**	**3.800**	**1.816**
LP &**V.O**	50	**5.881**	**1.860**	**0.492**	**5.895**	**1.288**	**0.045**
	100	**10.897**	**1.202**	**0.747**	**10.899**	**0.722**	**0.090**
	200	**20.897**	**2.027**	**1.338**	**20.898**	**0.912**	**0.190**
	300	**30.891**	**3.339**	**1.945**	**30.891**	**1.388**	**0.302**
	500	**50.85**	**5.550**	**3.560**	**50.85**	**1.900**	**0.608**
	1000	**100.9**	**15.700**	**7.578**	**100.9**	**3.700**	**1.546**

Table 2 display the results of IDRGP.MVE and DRGP.RMVN when $\alpha = 0.10$ for all outliers values (**LP**, **V.OUT**, LP &V.OUT) to 1000 samples . we noticed out of the results showed in this table that our suggested method is accurate and efficient and the reduction of the swamp when the size of the sample is less than 300 on.

Table -3-averages of the **correct** diagnosis, S**wap** and the time of computation, respectively, for three cases of simulation when $\alpha = 0.15$ and different sample sizes.

contamination	**n**	**DRGP(MVE)**			**DRGP(RMVN)**		
		correct	**Swap**	**Time**	**correct**	**Swap**	**time**
LP	50	**8.504**	**2.359**	**0.487**	**8.669**	**0.822**	**0.045**
	100	**15.695**	**1.474**	**0.742**	**15.707**	**0.334**	**0.089**
	200	**30.599**	**1.633**	**1.316**	**30.644**	**0.456**	**0.184**
	300	**45.543**	**1.985**	**1.915**	**45.591**	**0.579**	**0.291**
	500	**75.7**	**2.300**	**3.431**	**75.7**	**1.001**	**0.589**
	1000	**149.5**	**5.001**	**7.560**	**150**	**2.001**	**1.538**
V.OUT	50	**8.025**	**3.049**	**0.493**	**7.912**	**0.790**	**0.046**
	100	**15.632**	**2.471**	**0.752**	**15.761**	**0.341**	**0.093**
	200	**30.809**	**6.515**	**1.343**	**30.747**	**0.461**	**0.197**
	300	**45.794**	**10.485**	**1.974**	**45.708**	**0.598**	**0.318**

	500	**75.800**	**19.850**	**3.517**	**75.500**	**1.001**	**0.630**
	1000	**150.8**	**32.800**	**8.062**	**150.4**	**0.900**	**1.770**
LP &Y	50	**8.751**	**1.116**	**0.490**	**8.796**	**0.732**	**0.045**
	100	**15.837**	**0.580**	**0.744**	**15.847**	**0.396**	**0.089**
	200	**30.855**	**0.941**	**1.314**	**30.855**	**0.434**	**0.184**
	300	**45.835**	**1.395**	**1.932**	**45.847**	**0.613**	**0.295**
	500	**75.8**	**3.200**	**3.431**	**75.8**	**0.600**	**0.565**
	1000	**150.8**	**5.800**	**7.446**	**150.8**	**1.400**	**1.484**

The results showed in table (3) that the efficiency of IDRGP(RMVN) in the accuracy of diagnosis and fast measurement and reduction of the average of the incorrect diagnose when the pollution average is increased to 0.15 , it is not different from the efficiency and the supremacy when the average is 0.10 in table (2).

6. The Market value of Banks Iraq's Stock Market

The researchers collected these data out of the official website of the Iraqi Stock Markit after using the (SX60) system, where the annual data for market value were collected for nine of the local banks: Ashur International Bank For Investment, TBI Bank, Gulf Commercial Bank, Iraqi Middle East Investment Bank, Mousil Bank For Development& Investment, Babylon Bank, Bank Of Baghdad, Dijlah & Furat Bank for Development and Investment Bank of Iraq.

These banks were chosen due to it the most traded than others for the period (2011-2015). The data are contained eight variables and they are (Trading Rate, Earning per share (EPS), share turn over ratio, Annual Average price, the Assets, Undistributed earnings, Annual Net Profit (Revenue), and market value). The researchers are considered seven out of those variables explain and show the size of the market value according to the multiple linear regression model that can be described as follows:

$$y = \beta_\circ + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

where

y: represents: - the market value

x_1: represents: - trading rate

x_2 represents: - ESP

x_3 represents: - represents share turnover ratio

x_4 represents: - annual average price rate

x_5 represents: - the assets

x_6 represents: - Undistributed earnings

x_7 represents: - annual net profit

The results of the comparison between IDRGP(MVE) and IDRGP(RMVN) methods are displayed in Table (4) that is shown there is identical similarity in diagnosing (10) outliers being diagnosed correct diagnosis, and there are (2) clean values diagnosed as being outliers in IDRGP(MVE) method and there are no masking cases.

Table (4) shows masking and swamping method to (IDRGP.MVE) and IDRGP(RMVN) for market value data

Measure	Total	IDRGP.RMVN		
		Swamping	Masking	correct
IDRGP.MVE	12	2	0	10
IDRGP.RMVN	10	0	0	10

table

(4) shows the results of the diagnosis displayed in Figure (1) that there is a great closeness between the two sub shapes. So, we noticed that IDRGP(RMVN) detects 12 outliers but two them are swamping without masking cases and 10 outliers are matched with IDRGP.RMVN.

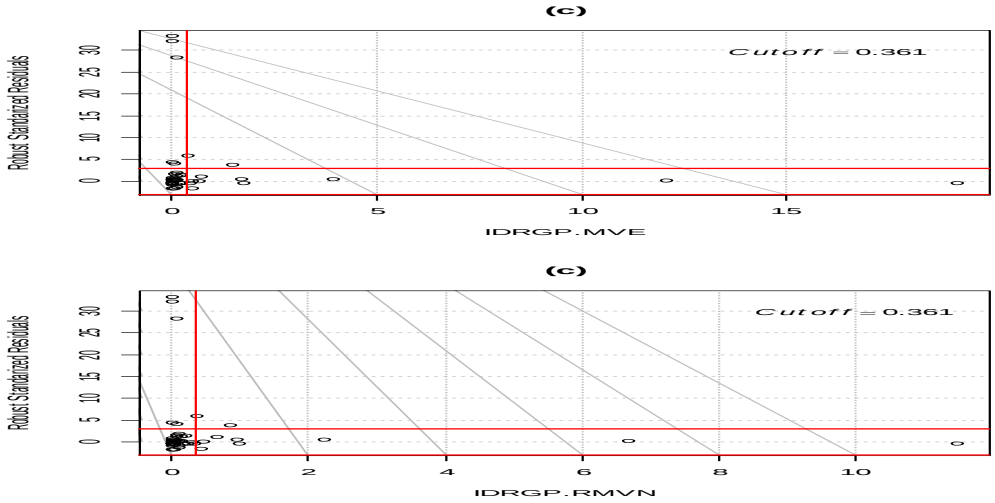

Figure (1) showed IDRG(MVE) and IDRG (RMVN) for the data of Market value data

7. Conclusion

This paper presented some of the diagnostic measures of the outlying observations in multivariate and regression data. In spite of DRGP(MVE) and IDRGP(MVE) have been shown distinct efficiency in group detection of leverage points, but both have not got rid of the effect of masking and swamping phenomenon perfectly. This problem motivated us to incorporate the RMVN estimator with IDRGP instead of the MVE estimator. That is because the working mechanism of the RMVN algorithm is based on repeated diagnostic within two concentration algorithms (DGK and MD), and reweighted (FCH) and (MVN) estimators two times to produce robust and scale estimators.

The efficiency of the new suggested method IDRGP(RMVN) was tested by comparison with the other method, through subjecting it to a number of simulation studies with different sample sizes with varying rates of pollution and for all types of

outliers, In addition to the test of its efficiency on the real data. We can conclude for simulation results that our suggested method showed stability and consistency to identify the correct outlying observations and reducing the rates of swamping cases. this accuracy of diagnostic lead to no masking cases unlike the IDRGP (MVE). Consequently, we recommend usage IDRGP(RMVN) for group diagnostic or group deletion measure of outliers for regression and multivariate data.

References

1. Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York :Wiley.
2. Devlin, Susan J, Gnanadesikan, Ramanathan, & Kettenring, Jon R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association, 76*(374), 354-362.
3. *Geary*, R. C. (1947). Testing for normality. *Biometriku 34,* 209-242. [1.3a].
4. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*. New York: Wiley.
5. Huber, P. J., & Ronchetti, E. M. (1981). Robust statistics, ser. *Wiley Series in Probability and Mathematical Statistics. New York, NY, USA, Wiley-IEEE*, *52*, 54.
6. Hadi, A. S. (1992). A new measure of overall potential influence in linear regression.*Computational and Statistical Data Analysis*. 14:1-27.
7. Imon, A.H.M.R. (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies*. *Special Volume in Honour of Professor Mir Masoom Ali*. 3: 207–218.
8. Midi, H., Ramli, N and Imon, A.H.M.R. (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): 507-520.
9. Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of theAmerican Statistical Association.* 79: 871–880.

10. Rousseeuw P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics.* 41:212–223
11. Mohammad A. Mohammad (2015), Rbust Estimation Methods and Robust Multicollinearity Diagnostics For Multiple Regression Model in The Presence of High Leverage Collinearity -Influential Observation, Thesis submitted to the School of Graduate Studies, UPM
12. Olive, David J, & Hawkins, Douglas M. (2010). Robust multivariate location and dispersion. Preprint, see (www. math. siu. edu/olive/preprints. htm).
13. Olive, David J. (2004). A resistant estimator of multivariate location and dispersion. *Computational statistics & data analysis*, *46*(1), 93-102.
14. Peirce, B., 1852. Criterion for the rejection of doubtful parametric and also moment free. Some of the procedure's observations. The Astronomical Journal, 2: 161-163