Group Diagnostic Measure of High Leverage Points

Assist. Prof. Hassan S. Uraibi[(1)] **Sawsan Abdul Ameer Haraj**[(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 topic of detection outliers is one of the important topics that have been of interest to researchers in many scientific fields. The presence of outliers in the dataset leads to the breakdown of the estimator of the method in use. The statistical literature has been shown that there are several types of outliers that occur according to the type and nature of the data, and therefore the researchers given concentrated on identifying is on the type of outliers of statistical models by using two diagnostic procedures, individual and grouped. Unfortunately, the individual procedure neglects the effect of the phenomenon of (masking and swamping), while the second procedure has not able to eliminate this phenomenon perfectly, but rather reduce the rates of its appearance. This paper seeks to suggest improvement of one of the well-known group diagnostic methods which so-called (DRGP) through the use of an RMVN location and scale matrix instead of MVE in order to reduce the effect of (swamping). The performance of a newly proposed method which denoted as (DRGP.RMVN) is tested with a number of simulation studies and real data. The results have shown that the performance of our proposed method is more efficient than (DRGP.MVE) to reduce the swamping points where the sample size is large in the presence of all types of outliers.

Keywords: Masking, Swamping, Leverage Point, DRGP and RMVN

Introduction

The topic of outlier’s detection in the samples data taken out of its statistical populations was not a topic that interest the researchers in diverse scientific fields until the sixties of the last century. It also was a reason that statistical school were divided into two schools, classical and robust. The classical school is stick to the theoretical bases to the assumption of the normal distribution of sample data which are drawn randomly from its statistical population. The founder Gauss had put a certain hypothesis that observations that are randomly chosen from its statistical population are independent and identically distributed. Most of the researchers found that one of the most important reasons behind the deviation of the normal distribution hypothesis is the presence of outliers, so it is of importance in the place of diagnosing these values that are considered far away from the center of the gathering bulk of data. Rousseuw and Zomeren (1990) defined the outliers as being observations that lie away from most of the rest data and it constitutes (1%) to (10%) out of any group of data in our real world. Recently, a group of researchers showed that this ratio could 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 (981) pointed out that the presence of one outlier at least in the data group leads to the breakdown of the statistical estimator. Great efforts were made in the statistical literature to diagnose all the types of outliers in linear regression such as individual diagnostic methods. Unfortunately, those methods did not take into consideration the phenomenon of masking and swamping which leads to their being unable to diagnose accurately for all types of outliers (Vertical Outliers (V.O) and High Leverage Point (HLP)), in the data set.

The individual diagnostic conceals in its 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. Consequently, Imon (2002) introduced group deleted measure as a Generalize Potential (GP) measure to get rid the effect of masking and swamping. Midi et al. (2009) found out that GP could not identifying the exact number of leverage points and still suffering from the effect of masking and swamping, therefore, 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 (DRGP). The target of algorithm is to the sake of accurate diagnostic and reducing the effect of masking and swamping. We noted that DRGP.MVE may tackle 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. Khan et al. (2007a), Khan et al. (2007b), Uraibi and Midi (2019) pointed out that MVE is a time-consuming procedure, even with the Fast algorithm of it that suggested by (Rousseeuw and Van Driessen, 1999), so it is not feasible option particularly with high dimensional data. 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 may RMVN is more relevant to DRGP than MVE. It is well known that DRGP.MVE algorithm is rely on Robust Mahanalobis Distance (RMD) that integrated with MVE estimators. In this paper, a slight development to the DRGP is proposed and we call it DRGP.RMVN by incorporating RMVN with RMD instead of MVE. This paper is organized to present the DRGP(MVE) Measure in Section 2. The Section 3 describes the

DRGP(RMVN) method. Section 4 and Section 5 illustrate simulation study and numerical example to assess the performance of the DRGP(RMVN) method.

1.2 DRGP(MVE) Measure

The idea of this method essentially relies on the first step in which robust-generalized diagnostics procedure for HLP by using MD with MVE location and scatter estimators. and then utilizing the GP algorithm proposed by Imon (2002). The algorithm of DRGP.MVE can be described as follows,

1. Computing the location $\hat{\mu}$ and scale $C_{MVE}(x)$ estimators of MVE.
2. Finding the mahalanobis distance MD by Eq. (1) and the i^{th} $MD \quad (MVE) > \sqrt{\chi^2_{(p,0.95)}}$ then the i^{th} row is having the suspected observations as HLP.

$$RMD_i(MVE) = \sqrt{[x-\hat{\mu}(x)]'[C_{MVE}(x)]^{-1}[x-\hat{\mu}(x)]} \; i = 1,2,\dots,n \quad (1)$$

3. The rows are determined including HLP's will delete from the design matrix x and put is in new submatrix denoted as X_D. The remaining rows that are having only clean observations will put in X_R matrix.
4. Constructing weight matrix as follows,

$$\mathrm{w} = \begin{bmatrix} \mathrm{U_R} & \mathrm{V} \\ \mathrm{V'} & \mathrm{U_D} \end{bmatrix}$$

where $U_\mathrm{R} = x_\mathrm{R}\left(x'x\right)^{-1} x'_\mathrm{R}$, $\mathrm{U_D} = x_\mathrm{D}\left(x'x\right)^{-1} x'_\mathrm{D}$, and

$V = x_R\left(x'x\right)^{-1} x'_D$.

when the D rows are omitted , the $W_{ii}^{(-D)}$is the i^{th} diagonal elements of $x'_i\left(x'_R x_R\right)^{-1} x_i \, , i = 1,2,,\dots,n,$ and $R = (N-D) \times p.$

Deletion the i^{th} diagonal elements from x_R makes $R = (N-1) \times p$, in this case $W_{ii}^{(-i)}$ will be a single diagnostic procedure equivalents to Hadi potential measure,

$$W_{ii}^{(-i)} = x'_i\left(x'_{(i)} x_{(i)}\right)^{-1} x_i = p_i$$

Finally, the group deletion measure based on MVE can be written as follows, $P_{\mathrm{ii}} = \begin{cases} W_{\mathrm{ii}}^{(-\mathrm{D})} & \forall \mathrm{i} \in \mathrm{D} \\ \frac{W_{\mathrm{ii}}^{(-\mathrm{D})}}{1-W_{\mathrm{ii}}^{(-\mathrm{D})}} & \forall \mathrm{i} \in \mathrm{R} \end{cases}$

When $P_{\mathrm{ii}} > median(P_{\mathrm{ii}}) + cMAD(P_{\mathrm{ii}})$ is confirmed the i^{th} row having HLP.

2.2- The DRGP(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 could be convergence during 5 to 10-steps.

Suppose that $(T_{5,DGK}, C_{5,DGK})$ are the DGK estimators and $(T_{5,MB}, C_{5,MB})$

Are MB estimators, the FCH location and scale estimators can be obtained by

$$T_{FCH} = \left\{ \begin{matrix} T_{5,DGK} & if\sqrt{|C_{5,DGK}|} < \sqrt{|C_{5,MB}|} \\ T_{5,MB} & Otherwise \end{matrix} \right\}$$

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

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 $MD_i^2[(\ \mathrm{T_{FCH}}, \mathrm{C_{FCH}})\ ] \le \chi^2_{(p,0.975)}$,and let $q_1 = min\left\{\frac{(0.5\times0.975\times n)}{n_1}, 0.995\right\}$

So, the first standard reweighting of MVN data is,

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

The new estimators ($\mathrm{T_{FCH}}, C_{RMVN}^{(1)}$) are applied to n_2 cases with $MD_i^2\left[(\ \mathrm{T_{FCH}}, C_{RMVN}^{(1)})\ \right] \le \chi^2_{(p,0.975)}$, 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,

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

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

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

$$MD_i(RMVN) = \sqrt{(x - T_{RMVN}(x))' C_{RMVN}^{(2)}{}^{-1'}(x - T_{RMVN}(x))}$$

3. Deletion D rows from matrix X of original data, where $D = MD\left\{(RMVN) > \sqrt{\chi^2_{(p,0.95)}}\right\}$ is the rows index and then put the deletion rows in X_D matrix, while the remining rows will be in X_R matrix.
4. The last step is similar to the step 4 in DRGP(MVE).

3-Simulation Study

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

$$y = x\beta + e \qquad (11)$$

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, s are 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 λ_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, **swap** and the time of computation, respectively, for three cases of simulation when $\alpha = 0.05$ and different sample sizes.

contamination	**n**	**DRGP(MVE)**				**DRGP(RMVN)**			
		E	**correct**	**swap**	**time**	E	**correct**	**swap**	**time**
LP	50	5.98	3.89	2.08	0.10	6.33	3.92	2.41	0.08

	100	7.22	5.92	1.31	0.18	7.79	5.92	1.87	0.11
	200	12.63	10.91	1.73	0.34	12.70	10.91	1.79	0.19
	300	18.20	15.90	2.30	0.53	18.14	15.90	2.24	0.26
	500	29.42	25.88	3.54	0.94	28.92	25.88	3.04	0.41
	1000	59.00	50.80	8.20	2.30	58.40	50.80	7.60	0.80
V.O.	50	5.97	3.81	2.16	0.10	6.50	3.86	2.64	0.08
	100	7.22	5.92	1.30	0.18	8.16	5.94	2.22	0.12
	200	12.62	10.93	1.69	0.35	12.73	10.93	1.80	0.19
	300	18.30	15.93	2.37	0.53	18.26	15.93	2.34	0.26
	500	29.60	26.00	3.60	0.98	29.50	26.00	3.50	0.41
	1000	57.60	50.50	7.10	2.39	57.60	50.80	6.80	0.80
LP &V.O.	50	5.93	3.94	1.99	0.10	6.63	3.94	2.69	0.08
	100	7.29	5.96	1.33	0.18	7.84	5.96	1.88	0.11
	200	12.64	10.96	1.68	0.35	12.69	10.96	1.74	0.19
	300	18.28	15.96	2.32	0.53	18.21	15.96	2.25	0.26
	500	28.20	26.00	2.20	0.95	28.01	26.00	2.01	0.41
	1000	57.10	50.90	6.20	2.30	56.70	50.90	5.80	0.81

Table -2- averages of the **correct** diagnosis, **swap** and the time of computation, respectively, for three cases of simulation when $\alpha = 0.1$ and different sample sizes.

contamination	**n**	**DRGP(MVE)**				**DRGP(RMVN)**			
		E	**correct**	**swap**	**time**	E	**correct**	**swap**	**time**
LP	50	7.35	5.78	1.58	0.10	7.39	5.83	1.56	0.08
	100	11.62	10.85	0.77	0.18	11.97	10.85	1.12	0.11
	200	21.70	20.78	0.92	0.34	21.73	20.79	0.94	0.18
	300	31.98	30.75	1.23	0.52	31.91	30.75	1.16	0.27
	500	52.70	50.90	1.80	1.04	52.60	50.90	1.70	0.44
	1000	102.80	100.10	2.70	2.32	102.80	100.20	2.60	0.82

V.OUT	50	6.44	4.85	1.60	0.10	7.22	5.14	2.09	0.08
	100	10.68	9.72	0.96	0.18	12.04	10.29	1.76	0.11
	200	20.54	19.51	1.03	0.35	21.07	20.07	1.00	0.19
	300	31.90	29.35	1.55	0.53	31.43	30.12	1.31	0.26
	500	95.90	94.00	1.90	0.98	95.51	94.01	1.50	0.41
	1000	104.60	99.80	4.80	2.52	104.30	100.40	3.90	0.85
LP & **V.O**	50	7.39	5.89	1.50	0.10	7.59	5.89	1.70	0.08
	100	11.72	10.90	0.82	0.18	11.94	10.90	1.04	0.11
	200	21.85	20.90	0.96	0.34	21.81	20.90	0.91	0.19
	300	32.16	30.89	1.27	0.52	32.08	30.89	1.19	0.26
	500	52.75	50.85	1.90	0.95	52.40	50.85	1.55	0.42
	1000	104.60	100.90	3.70	2.23	104.40	100.90	3.50	0.79

Table 1,2 and 3 display the results of DRGP.MVE and DRGP.RMVN when $\alpha = 0.05$,0.10,0.15 over all 1000 datasets generated with three types of contamination and different sample sizes $n = \{50,100,200.300,500,100\ \}$. It is obvious that the E cases of DRGP.MVE is less than the E cases of DRGP.RMVN when $(n = 50,100,200)$. On the other hand, the performance of DRGP.RMVN begins to be better than DRGP.MVE when $(n = 300,500,1000)$. The result shows the DRGP.RMVN is faster than DRGP.MVE. The results of Table 1,2 and 3 are shown, the DRGP.RMVN is more stable than DRGP.MVE for identifying the closet number to δ among E suspected cases in the presence of V.O. It is notable that the correct diagnostics of both methods are similar to each other when LP’s or both LP & V.O are present in the simulated data.

We noted that DRGP.RMVN methods is reduced the number incorrect diagnostics (swap) when the sample size is greater than or equals to 300 observations and it is not stable for sample sizes less than 300 observations. That means, DRGP.MVE has an ability to reduce (swap) with small sample sizes better that DRGP.MVE. Finally, the results of three tables expose that DRGP.RMVN is faster than DRGP.MVE.

Table -3- averages of the **correct** diagnosis, **swap** 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	9.47	**8.41**	**1.06**	**0.10**	9.83	**8.99**	**0.84**	**0.08**
	100	16.10	**15.69**	**0.41**	**0.18**	16.29	**15.71**	**0.58**	**0.11**
	200	31.10	**30.64**	**0.45**	**0.34**	31.08	**30.64**	**0.44**	**0.19**
	300	46.18	**45.59**	**0.59**	**0.51**	46.09	**45.59**	**0.50**	**0.26**
	500	76.70	**75.70**	**1.00**	**0.94**	76.60	**75.70**	**0.90**	**0.42**
	1000	152.00	**150.00**	**2.00**	**2.21**	151.90	**150.00**	**1.90**	**0.81**
V.OUT	50	5.21	**3.82**	**1.40**	**0.10**	6.22	**4.21**	**2.01**	**0.08**
	100	10.10	**9.28**	**0.82**	**0.18**	11.44	**9.83**	**1.61**	**0.11**
	200	20.27	**18.95**	**1.32**	**0.35**	20.30	**19.04**	**1.26**	**0.19**
	300	29.24	**27.45**	**1.79**	**0.54**	30.26	**28.62**	**1.64**	**0.26**
	500	42.40	**38.65**	**3.75**	**0.99**	48.20	**45.30**	**2.90**	**0.43**
	1000	119.50	**110.50**	**9.00**	**2.45**	155.00	**150.90**	**4.10**	**0.82**
LP &Y	50	9.66	**8.76**	**0.90**	**0.10**	9.56	**8.83**	**0.73**	**0.08**
	100	16.29	**15.85**	**0.44**	**0.18**	16.40	**15.85**	**0.55**	**0.11**
	200	31.32	**30.86**	**0.46**	**0.33**	31.30	**30.86**	**0.45**	**0.19**
	300	46.46	**45.85**	**0.61**	**0.51**	46.40	**45.85**	**0.55**	**0.26**
	500	76.40	**75.80**	**0.60**	**0.92**	76.20	**75.80**	**0.40**	**0.41**
	1000	152.30	**150.80**	**1.50**	**2.16**	152.10	**150.80**	**1.30**	**0.79**

3.1 Concrete Compressive Strength Data Set

In order to measure the efficiency of diagnosing the outliers in our suggested method we used the data of the compressive strength of concrete introduced by Yeh, 1998. This data includes (1030) observations collected from eight quantitative variables, Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, which are scaled by kg unit in a m^3 mixture. The ninth variable is Age scaled by day unit. The dependent variable is the Concrete compressive strength scaled by MPa.

In order to know the accuracy of the diagnosis, detection of the numbers of misdiagnosis (swamping) and the failure to detect some outliers (masking), we sought to make a comparison the traditional detection method, such as Hat Matrix, RMD, Hadi potential measure and DRGP.MVE. (Yeh) employed these variables to predict the strength of concrete compression according to a multiple linear regression model.

Table 4 explains the accuracy of the correct diagnostic and the incorrect diagnostic (swamping) and the failure in diagnosing the correct diagnostics (masking) for (Hadi, MD, hat) methods that compared with DRGP.MVE and DRGP.RMVN, and then compared DRGP.MVE and DRGP.RMVN. The results of the traditional method against DRGP.MVE appear there are (83) observations have been detected as (107) LP. There are (11) clean observations that have been detected as LP but this diagnostic is not correct. in addition, there is (35) observation considered being clean using (Hat) method but it is (LP) in DRGP(MVE) method. From table (4) we noticed that DRGP(RMVN) diagnosed (104) as (LP) and in comparison, with the (Hat), that the incorrect diagnostic is reduced to (10) observations as it had been in the previous method and the masking is reduced to (31) and the correct diagnostic is raised from (72) to (73).

Measure	**Total**	**DRGP.MVE**			**DRGP.RMVN**		
		Swamping	**Masking**	**Correct**	**Swamping**	**Masking**	**Correct**
Hat	**83**	**11**	**35**	**72**	**10**	**31**	**73**
MD	**84**	**7**	**30**	**77**	**6**	**26**	**78**
Hadi	**93**	**54**	**68**	**39**	**55**	**66**	**38**
DRGP.MVE	**107**				**5**	**2**	**102**
DRGP.RMVN	**104**						

Table (4) Diagnostic Masking, Swamping and Correct to Hadi, MD, Hat methods in comparison with DRGP(RMVN) and (DRGP(MVE) methods for the data of compressive strength of concrete.

As shown, the MD method diagnosed (84) LP that exceeds the cutoff value (5.71) which is a little less than what has diagnosed in the DRGP(MVE) method (see figure 1-a, 6-c) that detected (7) cases are being the wrong diagnostic, MD method recognised it as outliers, but are not, and (30) observations counted clean while they are outliers. For this reason, it is agreed upon the accuracy of diagnosing between the two methods on (77) observations only. The accuracy of diagnostic is raised to (78) in accordance with DRGP(RMVN) method that showed there are (6) clean observations being diagnosed as outliers in the MD method. On the other hand, this method neglected to diagnose (26) outliers.

Table (4) presented the results of the group diagnostics of (Hadi) method that was not completely successful compared to the previous two methods (Fig. (6)) neither in terms of the standard with our proposed method or DRGP (MVE) method or compared to Hat Matrix and MD method as it decreased. The correct diagnosis value was (39) compared to DRGP (MVE) in addition to the increase in the swamping values to (54) and masking to (68). As for the comparison with DRGP (RMVN), the value of the swamping increased to (55) and masking increased to (66), which indicates that this method is the worst performance among all methods.

Through Figure (1), which displays the diagnostic results mentioned in Table (4), we note that there is a very large convergence between Figure (1- (c)) and (6- (d)) related to the diagnostic results for the DRGP (MVE) and DRGP (RMVN) methods. Both methods agreed that there is (102) LP in the data of the compressive strength of the concrete and that our proposed method has identified that there are (5) cases that have considered as outliers by the DRGP (MVE) but which are not so, in addition to the diagnosis of two outliers that the DRGP (MVE) method could not detect .

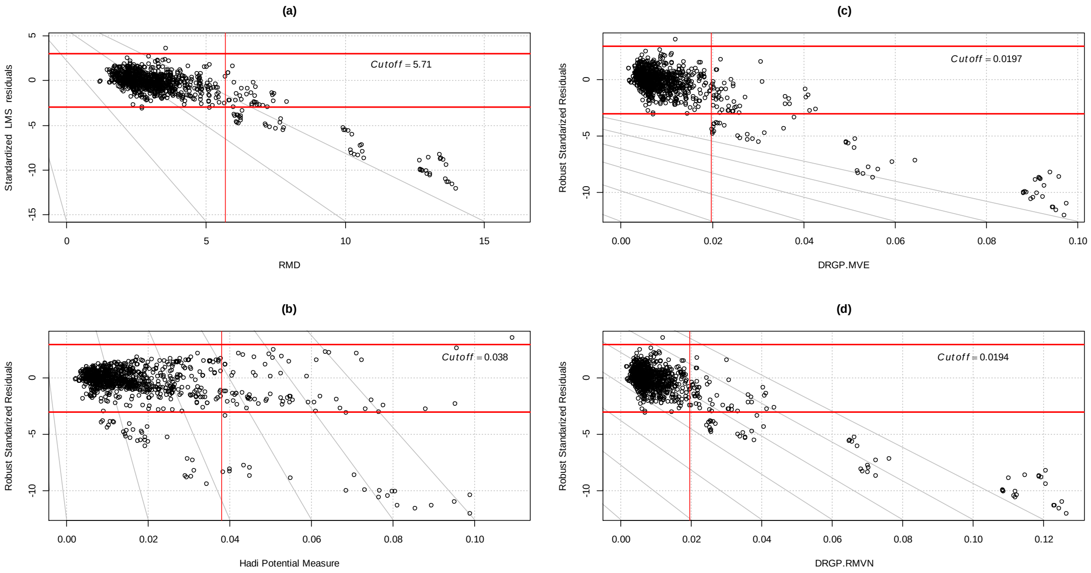

Figure (1) showed DRGP(RMVN) DRGP(MVE), RMD, Hadi) for the data of compressive strength of concrete

3.2 The Results

This research viewed some of the methods of the individual and group diagnostic to detect the outliers in the multivariable matrix and this is by using (Hadi Potential, RMD, Hat Matrix). These methods showed uneven efficiency in the accuracy of diagnostic especially with the presence of the two phenomena of swamping and masking. These shortcomings led to the developing of the idea of group diagnostic by some researchers like the DRGP(MVE) method that relies end on a robust variance and covariance matrix

(MVE). Unfortunately, MVE is time-consuming and combined with the DRGP algorithm was not tested it with large scale data. these reasons led us to substitute the MVE matrix with another one which is called (RMVN) and proposed a new method called DRGP(RMVN). The efficiency of our proposed method has been tested with the previous methods via subjecting it to a number of simulation studies that used samples of different sizes and different percentage of contaminating by using LP, V.O, HLP, LP&V.O. In addition to that testing its efficiency on real engineering data. We can conclude from the simulation outcomes that our suggested method proved consistency and stability in the accuracy of diagnostic and the reduction of the average of the incorrect diagnostic that the previous methods suffered from when the sizes of the samples were 300 and more. We noticed that there is a big closeness in the correct diagnosis for almost all types of outliers between our suggested method and the DRGP(MVE) method, yet the last method showed suffering to the problem of masking and swamping. That led to outperforms our method proposed on all the methods that are competing to it in the limits that it working on which is the large sizes of the samples and the different rates of contamination. So, we recommend the practitioners of statistics and researchers in this field to use our suggested method in diagnosing multivariate outliers or that are apparent in multiple linear regression data.

References

[1] A.H.M.R. , Imon, Identifying multiple high leverage points in linear regression. Journal of Statistical Studies, Special Volume in Honour of Professor Mir Masoom Ali. **3**(2002), 207–218.

[3] Devlin, Susan J, Gnanadesikan, Ramanathan, & Kettenring, Jon R. . Robust estimation of dispersion matrices and principal components, J J. AM. STAT. ASSOC., **76** (1981), 354-362. 3

[5] F. R.Hampel, E. M. Ronchetti, P. Rousseeuw and W. A. Stahel, Robust Statistics,Wiley, New York,(1986). 1

[6] H.Midi,N. Ramli, A.H.M.RImon, The performance of Diagnostic-Robust Generalized Potentials for the identification of multiple high leverage points in linear regression, J. APPL STAT. **36**(2009): 507-520. 1

[7] H. Midi, J. Arasan, H. S. Uraibi, H. Hendi., Fast and Robust Diagnostic Technique for the Detection of High Leverage Points. P.J.S.T, **28.** (2020) 1

[8] H. S. Uraibi, H. Midi, On Robust Bivariate and Multivariate Correlation Coefficient , Economic Computation & EconomicCybernetics Studies & Research, **53**(2019), 2.

[9] H. S. Uraibi, H. Midi, S. Rana, Selective overview of forward selection in terms of robust correlations, Communicationsin Statistics - Simulation and Computation, **47**,(2017) 5479-5503.

[10] H. S. Uraibi, H. Midi, S. Rana, Robust Stability Best Subset Selection for Autocorrelated Data Based on Robust Location

and Dispersion Estimator, Journal of Probability and Statistics, **2015**(2015) , 8 pages.

[11] J. A. Khan, S. Van Aelst, R. H. Zamar, Building a robust linear model with forward selection and stepwise procedures.

Computational Statistics & Data Analysis, **52** (2007a), 239-248.

[12] J . A.Khan, S. Van Aelst, Zamar, Ruben H., Robust linear model selection based on least angle regression, J. AM. STAT.ASSOC, **102**,(2007b), 1289-1299.

[13] Olive, David J., A resistant estimator of multivariate location and dispersion, Computational statistics & data analysis,,**46**,(2004), 93-102. 3

[14] Olive, David J, & Hawkins, Douglas M., Robust multivariate location and dispersion. Preprint, (2010) (www. math.siu. edu/olive/preprints. htm).

[15] P.J. Huber, Robust statistics, Wiley,New York, (1981).

[16] P.j. Rousseeuw and A. M. Leroy, Robust Regression and Outlier etection,Wiley, New York, (1987).

[17] P. j. Rousseeuw and B. Van Zomeren, Unmasking multivariate outliers and leverage points, J. AM. STAT. ASSOC., **85**(1999), 633-639 1

[18] P. J., Rousseeuw, Least median of squares regression, J. AM. STAT. ASSOC., **79**(1984), 871–880. 1

[19] P. J. Rousseeuw and K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator. Technometrics. **41** (1999):212–223 1