MALAYSIAN JOURNAL OF SCIENCE
MJS
ISSN 1394-3065

GROUP DIAGNOSTIC MEASURES OF DIFFERENT TYPES OF OUTLIERS IN MULTIPLE LINEAR REGRESSION MODEL

Hassan S. Uraibi [1a*] and Sawsan Abdul Ameer Haraj [2a]

Abstract: The topic of detection outliers is one of the crucial topics that have been of interest to researchers in many scientific fields. The presence of outliers in the dataset may lead to the breakdown of the estimator of the method in use. The statistical literature has shown that several types of outliers occur according to the type and nature of the data. Therefore, the researchers concentrated on identifying the type of outliers of statistical models by using two diagnostic procedures, individual and grouped. Unfortunately, the first procedure neglects the effect of the phenomenon of (masking and swamping). In contrast, the second procedure has not been able to eliminate this phenomenon ideally but rather reduce the rates of its appearance. This paper seeks to suggest improving one of the well-known group diagnostic methods (DRGP) by using an RMVN location and scale matrix instead of MVE to reduce the effect of (swamping). A newly proposed method denoted as DRGP(RMVN) is tested with a simulation study 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.

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

1. Introduction

The topic of outlier detection in the samples data taken out of its statistical populations was not a topic that interested researchers in diverse scientific fields until the sixties of the last century. It also was a reason that statistical schools were divided into two schools, classical and robust. The classical school sticks to the theoretical basis to assume the normal distribution of sample data drawn randomly from its statistical population (Uraibi and Alhussieny, 2021). On the other hand, the founder Gauss had put a particular hypothesis that randomly chosen observations from its statistical population are independent and identically distributed (Huber, 1981). Most of the researchers found that one of the most important reasons behind the deviation of the specific 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 centre of the gathering bulk of data (Hample et al., 1986). Apart from that, Rousseuw and Zomeren (1990) defined the outliers as being observations that lie away from most of the remaining data, which constitutes (1%) to (10%) out of any group of data in our natural 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 (Uraibi and Alhussieny, 2021).

Moreover, Huber (1981) 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 identify all the outliers in linear regression, such as single diagnostic methods (see, Rousseeuw and Leroy, 1987). Unfortunately, those methods did not take into consideration the phenomenon of masking and swamping, which leads to their being unable to detect all types of outliers (Vertical Outliers (VO) and High Leverage Point (HLP)) accurately in the data set. The single 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) (see, Maroona and Yohai, 2006).

On the other hand, may these methods suffer from the masking phenomenon in which the detected outliers probably overshadow other outliers. Therefore, the particular diagnostic method could not detect the outliers masked by other outliers (Rousseuw and Zomeren,1990). Consequently, Imon (2002) introduced a group deleted measure as a Generalize Potential (GP) measure to eliminate the effect of masking and swamping. However, Midi et al. (2009) found out that GP could not identify the exact number of leverage points and still suffer from the effect of masking and swamping. Therefore, they proposed utilizing Minimum Volume Ellipsoid (MVE) (Rousseeuw, 1984) to build a new algorithm which is a so-called Diagnostic Robust Generalized

Authors information:
[a]Department of Statistics, College of Administration and Economics, University of Al-Qadisiya, IRAQ. E-mail: hassan.uraibi@qu.edu.iq[1], stat.post22@qu.edu.iq[2]

*Corresponding Author: hassan.uraibi@qu.edu.iq

Received: January 22, 2022
Accepted: July 15, 2022
Published: September 30, 2022

DOI: https://doi.org/10.22452/mjs.sp2022no1.4

Potential measure (DRGP). The target of an algorithm is to the sake of accurate diagnostic and reducing the effect of masking and swamping. We noted that DRGP based on MVE (DRGP.MVE) may tackle the problem of identifying the exact number of leverage points. Still, it is not adequately effective in reducing the number of masking and swamping or getting rid of its effects.

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 these aspects, RMVN is more relevant to DRGP than MVE. On the other hand, it is well known that DRGP.MVE algorithm relies on Robust Mahanalobis Distance (RMD) that is integrated with MVE estimators, see (Uraibi and Midi;2009). 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. Meanwhile, Section 3 describes the DRGP(RMVN) method. Lastly, Section 4 and Section 5 illustrate simulation study and numerical examples to assess the performance of the DRGP(RMVN) method.

2. DRGP Measure

The idea of this method essentially relies on the first step in which a robust-generalized diagnostics procedure for HLP by using MD is employed with MVE location and scatter estimators. Then, the GP algorithm proposed by Imon (2002) is utilized. Suppose that $\boldsymbol{X}$ is a matrix of multivariate random varaibles. The algorithm of DRGP.MVE can be described as follows:

1. Computing the location $\hat{\boldsymbol{\mu}}$ and scale $\boldsymbol{C_{MVE}(X)}$ estimators of MVE, denoted as.
2. Finding the mahalanobis distance $(\boldsymbol{MD})$ using Eq. (1) if the $\boldsymbol{i^{th}}$ $\boldsymbol{MD}$ $\quad(\boldsymbol{MVE}) > \sqrt{\chi^2_{(p,0.95)}}$. Then. the $\boldsymbol{i^{th}}$ row has the suspected observations as HLP.
$$\boldsymbol{RMD_i(MVE)} = \sqrt{[X-\hat{\mu}(X)]'[C_{MVE}(X)]^{-1}[X-\hat{\mu}(X)]} \;\; i = 1,2,\ldots,n \quad (1)$$
3. The rows are determined including HLPs, which are deleted from the design matrix $\boldsymbol{X}$ and placed as a new submatrix denoted as $\boldsymbol{X_D}$. The remaining rows that have only clean observations will be substituted as $\boldsymbol{X_R}$ matrix. In other word, $\mathbf{R}$ and $\mathbf{D}$ are sets of any arbitrary remaining and deleted cases, respectively. Hence, R consist of $(\boldsymbol{n}-\boldsymbol{d})$ cases after $\boldsymbol{d}$ cases in $\mathbf{D}$ are deleted, where $\boldsymbol{d} < (\boldsymbol{n}-\boldsymbol{p})$, $\boldsymbol{n}$ is the sample size and $\boldsymbol{p}$ is the number of variables.
4. Habshah et al. (2009) pointed out without loss of generality, those observations are assumed to be the last of d rows of X, such that the weight matrix, $\boldsymbol{H} = \boldsymbol{X(X^tX)^{-1}X^t}$ can be decomposed as follows:
$$\mathbf{w} = \begin{bmatrix} \mathbf{U_R} & \mathbf{V} \\ \mathbf{V'} & \mathbf{U_D} \end{bmatrix},$$
where $\boldsymbol{U}_{\mathbf{R}} = \boldsymbol{X}_{\mathbf{R}}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X}'_{\mathbf{R}}$, $\mathbf{U_D} = \boldsymbol{X}_{\mathbf{D}}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X}'_{\mathbf{D}}$, are symmetric matrices of $(\boldsymbol{n}-\boldsymbol{d})$ and $\boldsymbol{d}$ cases, respectively, and $\boldsymbol{V} = \boldsymbol{X_R}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X}'_{\mathbf{D}}$ be an $(\boldsymbol{n}-\boldsymbol{d}) \times \boldsymbol{d}$ matrix.

When a group of observations $\mathbf{D}$ is omitted, the $\boldsymbol{W_{ii}^{(-\mathbf{D})}} = \boldsymbol{X}'_{\boldsymbol{i}}\left(\boldsymbol{X}'_{\boldsymbol{R}}\boldsymbol{X_R}\right)^{-1}\boldsymbol{X_i}$. Deletion the $\boldsymbol{i^{th}}$ diagonal element where $\mathbf{D} = \boldsymbol{i}$ result in $\boldsymbol{W_{ii}^{(-i)}} = \boldsymbol{X}'_{\boldsymbol{i}}\left(\boldsymbol{X}'_{(\boldsymbol{i})}\boldsymbol{X}_{(\boldsymbol{i})}\right)^{-1}\boldsymbol{X_i}$, which is a single diagnostic procedure equivalent to Hadi potential measure.

Finally, the group deletion measure based on MVE can be written as follows,

$$\boldsymbol{P}_{\mathbf{ii}} = \begin{cases} \boldsymbol{W}_{\mathbf{ii}}^{(-\mathbf{D})} & \forall \mathbf{i} \in \mathbf{D}, \\ \dfrac{\boldsymbol{W}_{\mathbf{ii}}^{(-\mathbf{D})}}{\mathbf{1}-\boldsymbol{W}_{\mathbf{ii}}^{(-\mathbf{D})}} & \forall \mathbf{i} \in \mathbf{R}. \end{cases}$$

Moreover, when $\boldsymbol{P}_{\mathbf{ii}} > \boldsymbol{median}(\boldsymbol{P}_{\mathbf{ii}}) + \boldsymbol{cMAD}(\boldsymbol{P}_{\mathbf{ii}})$, it is confirmed the $\boldsymbol{i^{th}}$ row has an HLP.

2.1 The DRGP(RMVN) Measure

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

Suppose that $(\boldsymbol{T}_{5,\boldsymbol{DGK}}, \boldsymbol{C}_{5,\boldsymbol{DGK}})$ are the DGK estimators and $(\boldsymbol{T}_{5,\boldsymbol{MB}}, \boldsymbol{C}_{5,\boldsymbol{MB}})$ are the MB estimators, then the Fast Consistence and Hgih breakdown (FCH) location and scale estimators can be obtained by

$$\boldsymbol{T_{FCH}} = \begin{Bmatrix} \boldsymbol{T}_{5,\boldsymbol{DGK}} & \boldsymbol{if}\sqrt{|\boldsymbol{C}_{5,\boldsymbol{DGK}}|} < \sqrt{|\boldsymbol{C}_{5,\boldsymbol{MB}}|} \\ \boldsymbol{T}_{5,\boldsymbol{MB}} & \boldsymbol{Otherwise} \end{Bmatrix},$$

and

$$\boldsymbol{C_{FCH}} = \begin{Bmatrix} \dfrac{\boldsymbol{MED}\left(\boldsymbol{MD}_i^2\left(\left(\boldsymbol{T}_{5,\boldsymbol{DGK}}, \boldsymbol{C}_{5,\boldsymbol{DGK}}\right)\right)\right)}{\chi^2_{(p,0.5)}} \times \boldsymbol{C}_{5,\boldsymbol{DGK}}, & \boldsymbol{if}\sqrt{|\boldsymbol{C}_{5,\boldsymbol{DGK}}|} < \sqrt{|\boldsymbol{C}_{5,\boldsymbol{MB}}|} \\ \dfrac{\boldsymbol{MED}\left(\boldsymbol{MD}_i^2\left(\left(\boldsymbol{T}_{5,\boldsymbol{MB}}, \boldsymbol{C}_{5,\boldsymbol{MB}}\right)\right)\right)}{\chi^2_{(p,0.5)}} \times \boldsymbol{C}_{5,\boldsymbol{MB}} & \textbf{Otherwise} \end{Bmatrix}$$

where |■| stands for the determinant of scale matrix while $\boldsymbol{MD}$ is the traditional Mahalanobis Distance.

DOI: https://doi.org/10.22452/mjs.sp2022no1.4

Let $(\widehat{T}_1, \widehat{C}_1)$ 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 $q_1 = min\left\{\frac{(0.5\times0.975\times n)}{n_1}, 0.995\right\}$

Thus, the first standard reweighting of MVN data is given by

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

The new estimators $(T_{FCH}, C^{(1)}_{RMVN})$ are applied to n_2 case with

$$MD_i^2\left[(\ T_{FCH}, C^{(1)}_{RMVN})\right] \leq \chi^2_{(p,0.975)}.$$

Let $q_2 = min\left\{\frac{(0.5\times0.975\times n)}{n_2}, 0.995\right\}$, then the RMVN estimator can be found as follows

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

The algorithm of DRGP (RMVN) measure can be summarized as follows

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

$$MD_i(RMVN) = \sqrt{(X - T_{RMVN}(X))' {C^{(2)}_{RMVN}}^{-1'} (X - T_{RMVN}(X))}\ .$$

3. The deletion of D rows from the matrix X of the original data where

$$D = MD\left\{(RMVN) > \sqrt{\chi^2_{(p,0.95)}}\right\}$$

is the row's index by placing the deletion rows in X_D matrix, while the remaining rows will be in X_R matrix.
4. The last step is similar to step 4 in DRGP(MVE).

3. Simulation Study

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

$$y = X\beta + e, \qquad (2)$$

where X is $n \times p$ design matrix generated from a multivariate normal distribution with mean equals to zero and standard deviation equivalent to $\sigma = \rho^{|i-j|}$, implying $x \sim N(0, \rho^{|i-j|})$. Here, $p = 7$, n is the generated sample that will take a different number of observations, $n = \{45, 70, 90, 140\}$, β is the identity vector of this model

$$\beta = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{7x1}, \qquad (3)$$

and e is a random error term that is distributed generally with zero mean and two standard deviations. To make sure of the diagnosis efficient of comparative methods, we contaminate the simulated data with different proportions of outliers, $\alpha = (0.05, 0.10)$ as follows:

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

2- Contaminating both design matrix and random errors α BLP & Vertical Outliers (VO) in the presence of one HLP. The VOs are generated from a chi-square distribution with (10) degree freedom.

The main reason for including a single HLP in all cases of the simulation study is to consider the phenomenon of masking and swamping. Let λ_i be a random variable, where $i = 1, 2, \dots, n$, $O = \{\lambda_1, \dots, \lambda_\delta\}$ be the outlying observations, such that $(\delta = \alpha \times n)$, and α is the percentage of outlying observations, respectively. The clean observations would be $C = \{\lambda_{\delta+1}, \dots, \lambda_n\}$. Suppose that $\mathbf{E}_j$ is the total outlying cases detected by a specific diagnostic method, where $1 \leq j \leq \delta^*$, δ^* is then either $(\delta + b)$ or $(+b)$, such that h and b are integer numbers, $[0 \leq b < n]$ and $[0 \leq h < \delta]$. Consequently, $\lambda_b \in C$ and $\lambda_h \in O$ and we can conclude that the exact detection will happen when $(\delta^* = \delta)$ in which no swamping cases $(b = 0)$ nor masking issues $(h = 0)$ will occur. However, the particular method would have swapping cases where $(\delta^* > \delta)$ and masking where $(\delta < \delta - h)$. The performance of our proposed method is compared with another overall (1000) dataset for each simulation case. The best diagnostic method is the one that has an average of correct diagnostic closer to δ (accurate), a lower standard of b (swap).

Tables 1,2 and 3 display the results of the Hat matrix, RMD, Hadi's poteintial, DRGP.MVE and DRGP.RMVN when α={0.05,0.10,0.15 } for overall 5000 datasets are generated with two types of contamination and different samples size n={35,45,70,90,140}. The average of of ($\mathbf{E}$, correct and swap) which are the number of outlying cases (Leverage points) that identified by competiting methods, the correct number of outlying cases and the number of swamping cases, respectively. For instance, when $(n = 35, \alpha = 0.05)$, the generated dataset should be having two LP, and probably a high LP that is generated randomly be either good or bad. If it is good high LP almost should be one of two leverage points, otherwise, the total number of LP will be three. This

DOI: https://doi.org/10.22452/mjs.sp2022no1.4

Table 1. Averages of the **correct** and **swap** diagnosis, respectively, for three cases of simulation when $\alpha = 0.05$ and different sample sizes.

			Hat	RMD	Hadi	DRGP(MVE)	DRGP(RMVN)
35	LP	E	11.4682	9.964	4.9952	4.2268	4.1232
		correct	2.9428	2.9428	2.9428	2.9428	2.9428
		swap	8.5254	7.0212	2.0524	1.284	1.1804
	LP & VO	E	11.4436	9.9408	5.396	4.282	4.1192
		correct	2.943	2.943	2.943	2.943	2.943
		swap	8.5006	6.9978	2.453	1.339	1.1762
45	LP	E	10.42	9.13	6.68	5.11	4.92
		correct	3.94	3.94	3.94	3.94	3.94
		swap	6.48	5.19	2.74	1.17	0.98
	LP & VO	E	10.41	9.18	7.27	5.05	4.98
		correct	3.93	3.93	3.93	3.93	3.93
		swap	6.48	5.25	3.34	1.12	1.05
70	LP	E	10.202	9.155	9.856	6.29	6.262
		correct	4.941	4.214	4.935	4.942	4.942
		swap	5.261	4.941	4.921	1.348	1.32
	LP & VO	E	10.234	9.218	10.49	6.217	6.159
		correct	4.939	4.939	4.93	4.94	4.94
		swap	5.295	4.279	5.56	1.277	1.219
90	LP	E	11.58	10.47	12.51	7.48	7.49
		correct	5.94	5.93	5.85	5.94	5.94
		swap	5.64	4.53	6.67	1.54	1.55
	LP & VO	E	11.54	10.53	13.21	7.46	7.49
		correct	5.95	5.95	5.88	5.95	5.95
		swap	5.60	4.58	7.33	1.51	1.54
140	LP	E	16.087	14.614	18.303	10.134	0.163
		correct	7.938	7.938	7.698	7.948	7.948
		swap	8.149	6.676	10.605	2.186	2.215
	LP & VO	E	16.034	14.547	19.332	10.033	10.041
		correct	7.93	7.926	7.773	7.945	7.945
		swap	8.104	6.621	11.559	2.088	2.096

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

			Hat	**RMD**	**Hadi**	**DRGP(MVE)**	**DRGP(RMVN)**
35	**LP**	E	7.83	7.193	6.646	5.811	5.623
		correct	4.865	4.865	4.838	4.865	4.865
		swap	2.965	2.328	1.808	0.946	0.758
	LP & VO	E	7.771	7.166	7.24	5.719	5.648
		correct	4.884	4.883	4.853	4.886	4.886
		swap	2.887	2.283	2.387	0.833	0.762
45	**LP**	E	8.837	8.208	7.945	6.683	6.547
		correct	5.878	5.874	5.553	5.895	5.895
		swap	2.959	2.334	2.392	0.788	0.652
	LP & VO	E	8.857	8.172	8.911	6.682	6.606
		correct	5.87	5.868	5.627	5.886	5.886
		swap	2.987	2.304	3.284	0.796	0.72
70	**LP**	E	11.9668	11.1538	10.96	8.7822	8.7544
		correct	7.8528	7.8432	6.9832	7.9048	7.9048
		swap	4.114	3.3106	3.9768	0.8774	0.8496
	LP & VO	E	11.99	11.1872	12.515	8.746	8.7206
		correct	7.8606	7.8518	7.1938	7.907	7.907
		swap	4.1294	3.3354	5.3212	0.839	0.8136
90	**LP**	E	14.9222	13.9716	13.5374	10.8636	10.8602
		correct	9.8176	9.8018	8.3428	9.9016	9.9016
		swap	5.1046	4.1698	5.1946	0.962	0.9586
	LP & VO	E	14.9008	13.919	15.569	10.8842	10.8764
		correct	9.8174	9.7996	8.6946	9.8984	9.8984
		swap	5.0834	4.1194	6.8744	0.9858	0.978
140	**LP**	E	22.3434	20.9508	19.8194	16.2092	16.2104
		correct	14.7168	14.6814	11.6562	14.8992	14.8992
		swap	7.6266	6.2694	8.1632	1.31	1.3112
	LP & VO	E	22.3826	20.9838	19.8032	16.2218	16.23
		correct	14.7156	14.6826	11.6198	14.901	14.901
		swap	7.667	6.3012	8.1834	1.3208	1.329

DOI: https://doi.org/10.22452/mjs.sp2022no1.4

Table 3. Averages of the **correct** and **swap** diagnosis, respectively, for three cases of simulation when $\alpha = 0.15$ and different sample sizes.

			Hat	**RMD**	**Hadi**	**DRGP(MVE)**	**DRGP(RMVN)**
35	**LP**	E	9.086	8.5024	6.6246	7.3208	7.2726
		correct	6.75	6.73	5.3072	6.832	6.832
		swap	2.336	1.7724	1.3174	0.4888	0.4406
	LP & VO	E	9.1104	8.5514	7.2782	7.285	7.2544
		correct	6.7526	6.7288	5.4162	6.832	6.832
		swap	2.3578	1.8226	1.862	0.453	0.4224
45	**LP**	E	10.4724	9.8478	7.8152	8.3042	8.2728
		correct	7.7378	7.713	5.9242	7.8452	7.8452
		swap	2.7346	2.1348	1.891	0.459	0.4276
	LP & VO	E	10.497	9.8884	8.8176	8.339	8.2986
		correct	7.7394	7.7162	6.0872	7.8412	7.8412
		swap	2.7576	2.1722	2.7304	0.4978	0.4574
70	**LP**	E	15.3946	14.5858	10.2204	12.3012	12.3036
		correct	11.5568	11.4976	7.2486	11.8536	11.8536
		swap	3.8378	3.0882	2.9718	0.450	0.450
	LP & VO	E	15.3966	14.5954	12.2674	12.3214	12.3102
		correct	11.5374	11.4888	7.789	11.8376	11.8376
		swap	3.8592	3.1066	4.4784	0.4838	0.4726
90	**LP**	E	19.1892	18.2664	12.2388	15.357	15.352
		correct	14.4146	14.3474	8.408	14.8388	14.8388
		swap	4.7746	3.919	3.8308	0.5182	0.5132
	LP & VO	E	19.1708	18.2016	15.1412	15.3638	15.361
		correct	14.4296	14.353	9.2282	14.8454	14.8454
		swap	4.7412	3.8486	5.913	0.5184	0.5156
140	**LP**	E	28.4406	27.054	17.7994	22.593	22.5958
		correct	21.19	21.0848	11.4674	21.851	21.851
		swap	7.2506	5.9692	6.332	0.742	0.7448
	LP & VO	E	28.4262	27.0326	22.6872	22.5732	22.577
		correct	21.1834	21.0786	13.0092	21.8482	21.8482
		swap	7.2428	5.954	9.678	0.725	0.7288

DOI: https://doi.org/10.22452/mjs.sp2022no1.4

procedure has been done for all simulation scenarios. Each table has the results of both diagnostics single detection and group diagnostic methods. Therefore, the discussion of results would be taken the performance of single diagnostic methods first and then the dicussion the results of group diagnostic has been considered with some details.

Tables 1,2 and 3 display the results of the Hat matrix, RMD, Hadi's poteintial, DRGP.MVE and DRGP.RMVN when α={0.05,0.10,0.15 } for overall 5000 datasets are generated with two types of contamination and different samples size n={35,45,70,90,140}. The average of of ($\mathbf{E}$, correct and swap) which are the number of outlying cases (Leverage points) that identified by competiting methods, the correct number of outlying cases and the number of swamping cases, respectively. For instance, when ($\boldsymbol{n = 35, \alpha = 0.05}$), the generated dataset should be having two LP, and probably a high LP that is generated randomly be either good or bad. If it is good high LP almost should be one of two leverage points, otherwise, the total number of LP will be three. This procedure has been done for all simulation scenarios. Each table has the results of both diagnostics single detection and group diagnostic methods. Therefore, the discussion of results would be taken the performance of single diagnostic methods first and then the dicussion the results of group diagnostic has been considered with some details.

The results of single diagnostic methods (Hat,RMD, and Hadi) that presented in Table 1, Hadi's potential method has proved its ability accuracy diagnostic than Hat matrix and RMD. When ($\boldsymbol{n = 35, 45}$) the average numbers of $\boldsymbol{E}$ cases and swap of Hadi's potential method are less than Hat,RMD methods. In spite of that all the E cases of single diagnostic methods are involved the correct number of outlying cases, but Hadi's potential method reduced the swamping cases to the minimum . his superiority of Hadi's potential than other single diagnostic methods method has not held long. The signs of broken have started of this method is to be clear when ($\boldsymbol{n = 70}$) and there is vertical outliers and leverage points were present togather in the data. Table 1 has shown that RMD method is more accurate than Hadi's potential method when data are contaminated by LP & VO and ($\boldsymbol{n = 70, 90, 140}$) or in the presence of LP and ($\boldsymbol{n = 90, 140}$). We recorded that the single diagnostic methods may suffer from some masking cases particularly when the correct phases of it is more lower than their counterparts of group diagnostics. It is notable that Hadi's potential method started to be far from the correct cases gradually with the sample sizes are increased. The results that displays in Table 2 and 3 confirmed the outperforming the method of RMD than Hadi's potential and Hat matrix methods when $\{\boldsymbol{n = 45, 70, 90, 140}\}$ and where the outliers is presence in $\boldsymbol{n = 35}$ obsrvation. In another word, the Hadi's potential method are much influenced by masking cases than others as Table 2 and 3 are shown.

The performance of both group diagnostics methods DRGP.MVE and DRGP.RMVN are displayed in the Tabel 1,2, and 3. It is obvious that when 0.05, 0.10 of LP or LP and outliers together are present in the dataset, the total number of outlying cases (which is called $\boldsymbol{E}$ cases) that diagnostic by DRGP.RMVN method is less than the $\boldsymbol{E}$ cases of DRGP.MVE when ($\boldsymbol{n = 35, 45}$). However, Table 1 shows that $\boldsymbol{E}$ cases of five compared methods are 11.4682, 9.964, 4.995, 4,2268, and 4.1232, respectively. The closest number to (3) is 4.1232 which is determined by DRGP.RMVN method as the average of LP's that identified for overall 5000 iterations. The second method is DRGP.MVE which is detected 4.2268 LP's and Hadi's potential diagnosed 4.995. The Hat matrix and RMD methods are determined 11.4682 and 9.964 LP's, respectively. The good thing is that the E cases of all methods have been selected with the same number (2.9428) for the correct cases, but subtracting this number from the E cases of each method result-in the swamping cases.

Surely, the less number of (swap) will be the criterion for choosing the best method. Definitely, the results of Table 1 present that DRGP.RMVN is having a lower number of swap (1.1804) than others. in spite, of the swap of DRGP.MVE is very close to DRGP.RMVN, but the last method reduced the percentage of swap to 10%. The performance of all methods has been not changed in the second scenario of simulation (in the presence of five percent of outliers and leverage in the data) and outperforms DRGP.RMVN than DRGP.MVN and single detection methods even $\boldsymbol{n = 70}$ by two kinds from simulation scenarios are used. The DRGP.MVE method has proved its ability to compete with DRGP.RMVN method at (n=90,140) as Table 1 has been shown. That is Due to the values of the averages swamping DRGP.MVE is less than others.

The superior performance of DRGP.RMVN method has held even with increasing the sample size to (45,70) and the percentage of outlying observations increased to 10% as table 2 is showed that too. The performance of DRGP.MVE begins to get better than DRGP.RMVN when ($\boldsymbol{n = 90}$), ($\boldsymbol{\alpha = 0.05}$), but when $\boldsymbol{n = 90}$ with increasing $\boldsymbol{\alpha}$ to $\boldsymbol{0.10}$ DRGP.RMVN kept its high performance compared with other methods. Where ($\boldsymbol{n = 140}$), and ($\boldsymbol{\alpha = 0.05, 0.10}$) the DRGP.MVE shows its ability to identfy the LP's than others. But the outcomes of Table 3 confirmed the high diagnostics acuuracy of DRGP.RMVN than DRGP.MVE even though sometimes the performance of both methods are equevalant.

3.1 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. These banks were chosen due to it the most traded than others for the period (2011-2015). The 45 samples are contained eight variables and they are (Trading Rate $\boldsymbol{x_1}$, Earning per share (EPS) $\boldsymbol{x_2}$, share turn over ratio $\boldsymbol{x_3}$, Annual Average price $\boldsymbol{x_4}$, the Assets

DOI: https://doi.org/10.22452/mjs.sp2022no1.4

x_5, Undistributed earningsx_6, Annual Net Profit (Revenue) x_7, and market value y). We 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:

$$y = \beta_{\circ} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 \qquad (4)$$

Figure 1 shows that the distribution of the residuals of model (4) is not follow normal distribution. It is clear, in the Q-Q plot, the fitted value vs residuals and scale location appear the resfigureiduals which are indexed in 25, 29,and 30 are outliers. The plot of residuals vs leverage points recorded some leverage points the indentified by Cook's distance measure.

Table 4 explains the accuracy of the correct diagnostic and the incorrect diagnostic (swamping & masking) for (Hadi, MD, Hat) methods compared with DRGP.MVE and DRGP.RMVN methods. Moreover, the DRGP.MVE and DRGP.RMVN are compared to each other.

Based on the simulation results DRGP.MVE and DRGP.RMVN methods are high efficiency and more accuarate than single diagnostic method to detect the leverage points. So, we consider it as criterion to identifying the correct and non-correct diagnostic. There is (45) samples of Banks market values probably motivate us to expect that DRGP.RMVN is more stable and accuracy diagnostics than DRGP.MVE. That is due to the simulation result shows the high performance of DRGP.RMVN with the small samples. The DRGP.MVE and DRGP.RMVN are determined (10) and (9) samples which are (1, 6, 7, 16, 23, 33, 34, 35, 36, 40) and (1, 6, 7, 13, 16, 33, 34, 35, 40) having LP's, respectively.

The Hat matrix identifies (18) samples which are (12, 13,14,16, 18, 23, 24, 25, 27, 31, 32, 33, 34, 35, 37, 38, 39, 40) are having LP's. The comparison of Hat matrix result with DRGP.MVE method, we noted that both methods are only matched to identify (6) samples that poses correct leverage points which are (16, 23, 33, 34, 35,40), while (12) samples

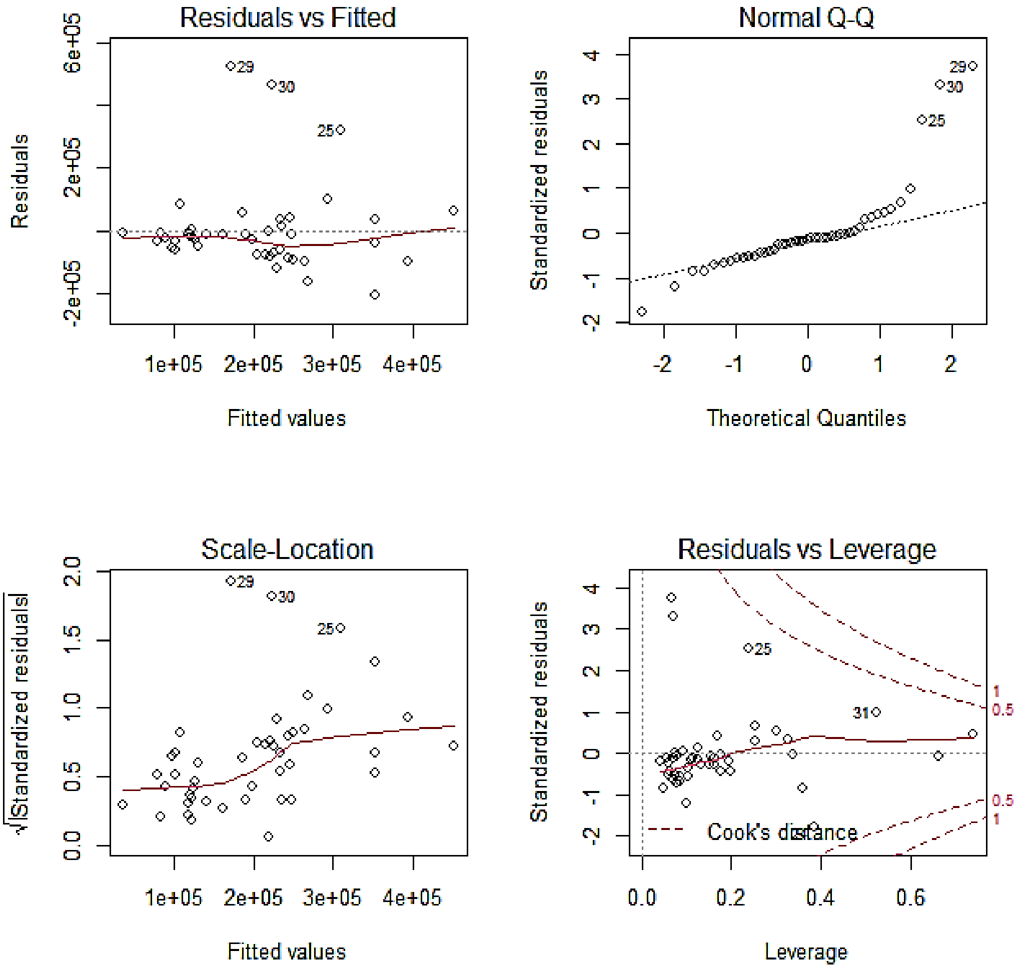

Figure 1. Initial diagnostics of outliers and leverage points for bank market value data

DOI: https://doi.org/10.22452/mjs.sp2022no1.4

Table 4. Diagnostic Masking, Swamping and Correct to Hadi, MD, Hat methods in comparison with DRGP(RMVN) and DRGP(MVE) methods in terms of market value data.

Measure	Total	DRGP.MVE			DRGP.RMVN		
		Swamping	Correct	Masking	Swamping	Correct	Masking
Hat	18	12	6	4	12	6	3
MD	13	3	10	0	4	9	0
Hadi	8	5	3	7	1	3	6
DRGP.MVE	10			0	2	8	1
DRGP.RMVN	9						

are considered leverage points by Hat matrix, but are not detected by DRGP.MVE. So, this wrong diagnostics is pointed as swamping cases. On the other hand, the DRGP.MVE recorded (1,6,7,36) samples involved leverage points that are not detected by Hat matrix, therefore, we considered these as masking cases in Hat matrix. The comparison of Hat matrix with DRGP.RMVN method has not differ a lot, just it is reduced the masking cases to (3).

The MD method has reduced the total detection of leverage points from (18) case with Hat matrix to (13) case which are (1, 6, 7, 11, 13, 16, 23, 32, 33, 34, 35, 36, 40). This procedure is already would reduce the swamping cases to (3) and (4) compared with DRGP.MVE and DRGP.RMVN, respectively. So, the correct diagnostic of MD method matches with the correct diagnostics of DRGP.MVE and DRGP.RMVN without any masking cases. Unfortunately, Hadi's potaintial method detect (8) LP"s that are noted in (24, 25, 29, 30, 31, 34, 35, 40) samples, but only (3) samples are matched with DRGP.MVE and DRGP.RMVN methods {34, 35, 40} and other are swamping cases. Due to the difference in total detection of leverage points between DRGP.MVE and DRGP.RMVN methods are only one case, therefore the masking cases of Hadi's potential method are (7) and (6) compared with both of the previous methods, see Table ( ). However, DRGP.RMVN method has found that (9) samples are contained leverage points without any swamping and masking cases. This outcome is compatible with the simulation scenario where $n = 45$.

Figure 2 contains (6) subgraphs, each graph shows the behavior of a certain diagnostic method against the standardized residuals measure. The vertical line represents the cutoff point of that diagnostic method, while the horizontal line represents the threshold of standardized residuals which equals 3 in this paper. It is obvious, that the developments that have happened in the detection methods of leverage points have reduced the swamping cases. The first subgraph of the Hat matrix method confirms that there are (18) leverage points and the second subgraph of MD displayed only (2) swamping cases. The third subgraph of RMD presents the high performance of RMD vs MD and has reduced the swamping cases better than Hat matrix method.

Figure 2 contains (6) subgraphs, each graph shows the behavior of a certain diagnostic method against the standardized residuals measure. The vertical line represents the cutoff point of that diagnostic method, while the horizontal line represents the threshold of standardized residuals which equals 3 in this paper. It is obvious, that the developments that have happened in the detection methods of leverage points have reduced the swamping cases. The first subgraph of the Hat matrix method confirms that there are (18) leverage points and the second subgraph of MD displayed only (2) swamping cases. The third subgraph of RMD presents the high performance of RMD vs. MD and has reduced the swamping cases better than the Hat matrix method.

Hadi's potential method displayed in the subgraph fourth could not deal with these specific cases, therefore, we noted that it identified some cases that are not detected by other methods. Finally, the fifth and sixth subgraphs have related to DRGP.MVE and DRGP.RMVN methods and due to their asymptotic performances to each other, both graphs seem to be similar, but in reality, are a little bit different.

3.2 The Results

This research viewed some individual and group diagnostic methods to detect the outliers in the multivariable matrix using (Hadi Potential, RMD, Hat Matrix). However, these methods showed uneven efficiency in diagnostic accuracy, especially with the presence of the two phenomena of swamping and masking. These shortcomings led to the development of group diagnostic by some researchers like the DRGP.MVE method that relies on a robust variance and covariance matrix (MVE). Unfortunately, MVE is suffering from swamping cases, particularly with small samples. This reason led us to substitute the MVE matrix with another one called (RMVN) and proposed a new method called DRGP(RMVN). The efficiency of our proposed

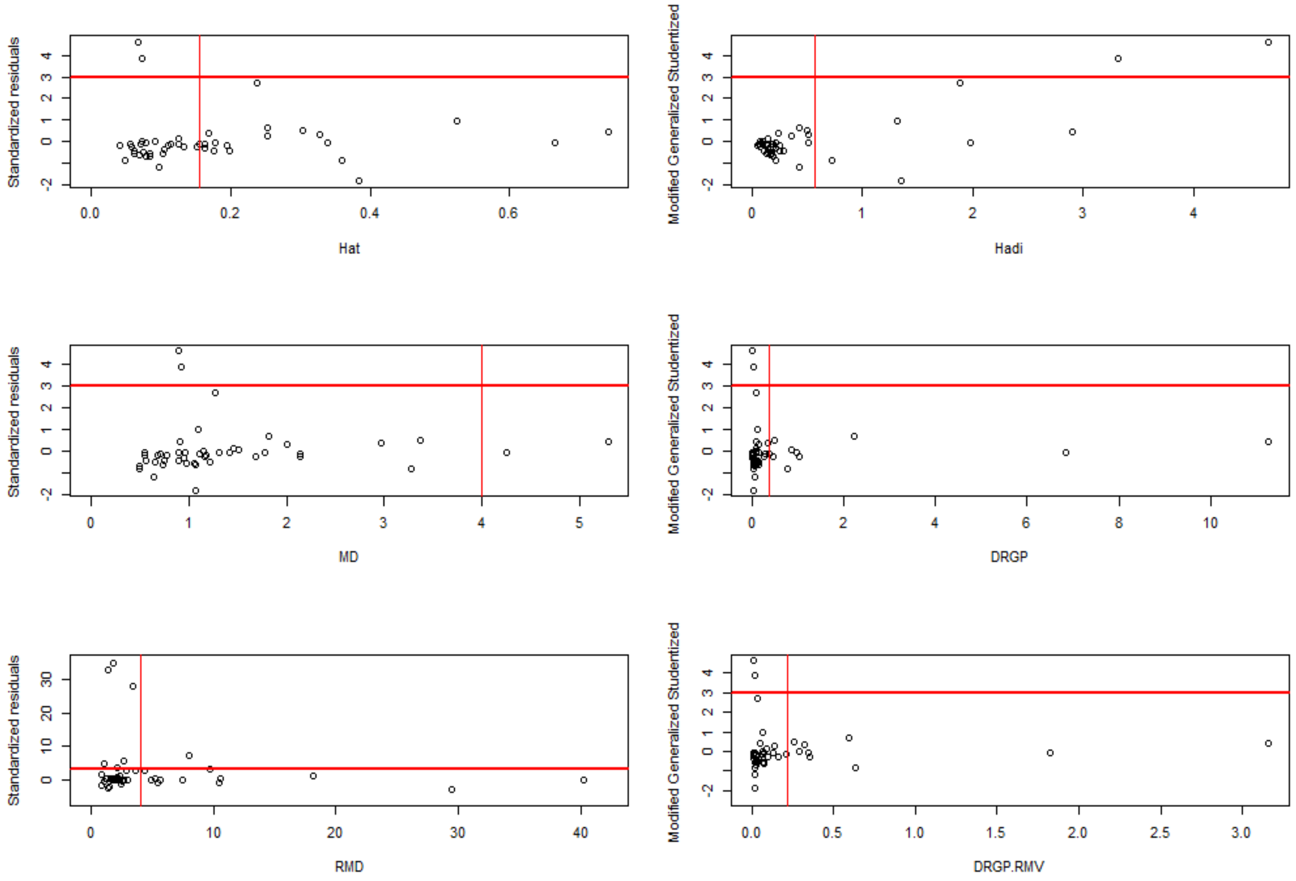

Figure 2. DRGP(RMVN), DRGP(MVE), RMD, Hadi for the banks market value data of Iraq's Stock Market

method has been tested with the previous techniques by subjecting it to many simulation studies using different sizes' samples and contaminating's different percentages of LP and (LP & VO). This is in addition to testing its efficiency on actual finance 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 techniques suffered from when the sizes of the samples were 35,45, and 70.

Furthermore, we noticed an enormous closeness in the correct diagnosis for LP's between our suggested method and the DRGP.MVE approach. Yet, the final form showed suffering in the problem of masking and swamping. That led to outperforming our proposed method among all the methods competing within limits, for example, the small sizes of the samples and the different rates of contamination. Thus, we recommend that the practitioners of statistics and researchers in this field use our suggested method to diagnose multivariate outliers apparent in multiple linear regression data.

4. References

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.

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.

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

H.Midi, N. Ramli, A.H.M.RImon, The performance of Diagnostic-Robust Generalized Potentials to identify multiple high leverage points in linear regression, J. APPL STAT. 36(2009): 507-520.

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

H. S. Uraibi, S. A. Alhussieny, Improvise Group Diagnostic Potential Measure for Multivariate Normal Data, Al-

Qadisiyah Journal for Administrative and Economic Sciences, 23,(2021),2.

I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998).

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

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

P.J. Huber, Robust Statistics, Wiley, New York, (1981).

P.j. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, (1987).

P. j. Rousseeuw and B. Van Zomeren, Unmasking multivariate outliers and leverage points, J. Am. STAT. ASSOC., 85(1990), 633-639

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

R.A. Maronna, R. D. Martin and V.J. Yohai, Robust Statistics Theory and Methods. New York: Willy and sons, (2006).