Sains Malaysiana 50(3)(2021): 859-867
http://dx.doi.org/10.17576/jsm-2021-5003-26

Simple and Fast Generalized - M (GM) Estimator and Its Application to Real Data Set

(Penganggar Ringkas dan Pantas Teritlak- M dan Kegunaannya ke atas Set Data Sebenar)

HABSHAH MIDI*, SHELAN SAIED ISMAEEL, JAYANTHI ARASAN & MOHAMMED A MOHAMMED

ABSTRACT

It is now evident that some robust methods such as MM-estimator do not address the concept of bounded influence function, which means that their estimates still be affected by outliers in the X directions or high leverage points (HLPs), even though they have high efficiency and high breakdown point (BDP). The Generalized M(GM) estimator, such as the GM6 estimator is put forward with the main aim of making a bound for the influence of HLPs by some weight function. The limitation of GM6 is that it gives lower weight to both bad leverage points (BLPs) and good leverage points (GLPs) which make its efficiency decreases when more GLPs are present in a data set. Moreover, the GM6 takes longer computational time. In this paper, we develop a new version of GM-estimator which is based on simple and fast algorithm. The attractive feature of this method is that it only downs weights BLPs and vertical outliers (VOs) and increases its efficiency. The merit of our proposed GM estimator is studied by simulation study and well-known aircraft data set.

Keywords: DRGP; GM-estimator; high leverage points; index set equality

ABSTRAK

Beberapa kaedah teguh seperti penganggar MM telah dibuktikan tidak dapat menanangi konsep fungsi pengaruh terbatas, yang membawa maksud bahawa penganggar MM masih terjejas dengan titik terpencil dalam arah X atau dikenali sebagai titik tuasan tinggi (HLPs), walaupun ia mempunyai kecekapan dan titik musnah (BDP) yang tinggi. Penganggar -M teritlak (GM), seperti penganggar GM6 dicadangkan dengan tujuan utama membuat batasan kepada pengaruh HLPs dengan fungsi pemberat. Penganggar GM6 mempunyai kekangan dengan memberi pemberat rendah kepada GLPs, yang mengakibatkan kecekapan penganggar ini menurun apabila kehadiran HLPs bertambah banyak dalam suatu set data. Tambahan pula, masa pengiraan GM6 terlalu panjang. Dalam kertas ini, kami membangunkan penganggar GM versi baru berdasarkan algoritma yang mudah dan pantas. Sifat menarik yang ada bagi kaedah ini ialah ia hanya menurunkan pemberat bagi BLPs dan VOs dengan ini kecekapannya meningkat. Merit penganggar GM yang kami cadangkan telah dikaji melalui kajian simulasi dan set data kapal terbang yang terkenal.

Kata kunci: DRGP; penganggar GM; set indek kesamaan; titik musnah tinggi

INTRODUCTION

The ordinary least squares (OLS) is the widely used method in multiple linear regression due to tradition and its optimal properties. However, in the presence of outliers in a data, the OLS estimates become inefficient. Several versions of outliers are defined in regression problems such as residual outliers (ROs), high leverage points (HLPs) and vertical outliers (VOs). Any observation that has large residual is referred to as residual outlier. Vertical outliers are those observations that are extreme or outlying in y-coordinate. High leverage points (HLPs) not only fall far from the majority of independant variables, but also are deviated from a regression line because they actually tilt the OLS line and their effect on OLS estimator is very large (Leroy & Rousseeuw 1987). According to Midi et al. (2009), the detection of HLPs is very crucial due to its responsibility for misleading conclusion about the fitting of regression model, causing multicollinearity, and masking/swamping of outlier. Hence the effect of HLPs should be minimized to get more efficient estimate. Nonetheless not all high leverage points (good or bad) have an adverse effect on the OLS estimates. It is now evident that the OLS is only

affected by bad leverage points. According to Chatterjee and Hadi (2006), good leverage points contribute to the efficiency of an estimate since they follow the pattern of the majority of a data.

To remedy the problem of outliers on the parameter estimates, robust methods which are known to be resistant to outliers may be employed. Many robust estimation methods such as M, MM, LMS and LTS can be found in the literatures (Huber 2004; Leroy & Rousseeuw 1987; Yohai 1987; Wilcox 2005). Even though some of them have high efficiency and possess high breakdown point (HBDP), they do not have bounded influence properties (Simpson et al 1992). Yohai and Zamar (1988) pointed out that one of the aims of robust regression is to achieve high efficiency, high breakdown point (close to 50%), and bounded influence properties. The breakdown point of M estimator is very low which is equals to (1/*n*). It can handle vertical outliers but not successful in handling HLPs. Hekimoğlu and Erenoglu (2013) noted that both the S and MM estimators also do not have bounded influence property, despite of having high breakdown and high efficiency. On the other hand, both LTS and LMS have high breakdown point, but they do not have bounded influence property and have very low relative efficiency which is close to 8 and 37%, respectively (Rousseeuw 1984; Rousseeuw & Croux 1993; Stromberg et al. 2000). Since none of these estimators can handle high leverage point, Schweppes as described by Hill and Paul (1977) suggested a new robust method called bounded influence Generalized M-estimator (GM-estimator) as a remedial technique for the sensitivity of M-estimator against high leverage points (Andersen 2008; Hill & Paul 1977).

Many types of GM-estimators were proposed in literature (Andersen 2008; Wilcox 2005). However, these methods have achieved a moderate BDP equals to 1/*k*, where *k* is the number of regression coefficients including the intercept (Simpson et al. 1992). As a remedial measure, multi-stage GM-estimators were developed. The most popular types of multi-stage GM-estimator is GM6 which was introduced by Coakley and Hettmansperger (1993). The least trimmed of squares is employed as an initial estimator in the algorithm of GM6. The initial *d*–weight function of GM6 estimator is expressed in terms of robust mahalanobis distance (RMD) which utilized robust location and scatter estimators obtained from minimum volume ellipsoid (MVE) (Rousseeuw 1985). It is noted that MVE suffers from swamping effect and long computation running times. Besides, the RMD which is based on MVE only attempts to identify HLPs which may consist of GLPs and BLPs. Thus, the GM6 efficiency inclines to decrease as the number of GLPs increases because both GLPs and BLPs are down weighted. Their work has motivated us to develop another version of GM estimator which is relatively simple, easy to understand and fast.

In this paper, we propose another version of GM estimator that we call Generalized M estimator based on Fast Improvised Generalized MT (FIMGT) estimator denoted as GM-FIGMT which is quite fast and satisfy all the three properties of good robust method. The Fast GM estimator utilizing high breakdown point S-estimator as an initial estimate and using more effective weight function based on FIMGT. The merit of the FIMGT is that it can correctly identify VOs, GLPs and BLPs with relatively less computer time. The FIMGT is adapted in the formulation of the GM estimator whereby it only downs weight BLPs and VOs and assigns weight equals 1 to GLPs. The good leverage points are not down weighted because they may contribute to the precision of the estimates as their presence have no impact or less effect on the OLS estimates (Andersen 2008; Rousseeuw & Van Zomeren 1990).

This paper is organized as follows. The procedure for formulating the proposed initial weight function for the GM estimator is presented in the next section. The proposed procedure highlights the choice of the initial weight d_i. Subsequently, the proposed GM estimator is explained in detail. Monte Carlo simulation study and real aircraft data are illustrated in the following section. The last section summarizes the conclusion of the study.

Materials and Methods

The Proposed Initial Weight for GM-Estimator

In this section, the existing and the proposed initial weight functions used in the GM estimator are discussed. The choice of the weight function is based on the detection method of HLPs. A good initial weight function is one that depends on the detection method that able to correctly identify VOs and BLPs.

Choice of d_i Weight Function for GM6

Coakley and Hettmansperger (1993) introduced GM6 estimator which has high efficiency at normal distribution, bounded influence property and high breakdown point. It can be expressed as a solution of normal equations given by

$$\sum_{i=1}^{n} d_i \psi\left(\frac{y_i - x_i^t \hat{\beta})}{\hat{\sigma} d_i}\right) x_i = 0 \qquad (1)$$

where $\psi = \rho'$ is an influential function and $d_i = 1, 2, ..., n$ is the initial weight function.

The GM estimators' main objective is to downweight HLPs which have large residuals. Coakley and Hettmansperger (1993) employed RMD based on MVE or MCD, using $\chi^2_{(0.95,p)}$ as cut-off points. Those detected HLPs will be assigned smaller weight while regular observations are given weight equals 1.0.

Afterwards, they defined the initial weight of the GM6 estimator as follows:

$$d_i = \min\left[1, \left(\frac{\chi^2_{(0.95,p)}}{RMD^2}\right)\right], i = 1,2, \ldots, n$$

Bagheri and Midi (2015) noted that this initial weight function inclines to swamp some low leverage points. Another limitation of this weight function is that, the RMD only identify HLPs (good and bad). This implies that the detected HLPs will be assigned low weight irrespective of whether they are GLPs or BLPs. Thus, as the number of GLPs increases, the GM6 efficiency tends to decrease because the precision of the parameter estimates may be contributed by GLPs as noted by Rousseeuw and Van Zomeren (1990). This is the reason why the GM6 - estimate is less efficient because both GLPs and BLPs are downweighted. The computation of GM6 estimator is very long since it uses MVE or MCD. This contributes to another weakness of GM6 estimator.

THE PROPOSED INITIAL WEIGHT FOR THE NEW PROPOSED GM ESTIMATOR

Our propose GM estimator begins by establishing an algorithm of detecting VOs, GLPs and BLPs at the outset. Thereafter, only assign smaller weights to the detected VOs and BLPs and weights equal 1 to GLPs, to increase the efficiency of the GM estimator. A simple and fast method is incorporated in the establishment of the algorithm of classification of observations into VOs, GLPs and BLPs. To make our method fast, we employ Index Set Equality (ISE) to compute the location and scale estimators instead of using the MVE or MCD. The computation time of ISE is shown to be quicker than the MVE or even quicker than the fast MCD (Lim & Midi 2016). Hence, our aim is to formulate an initial weight whereby only minimize the detected VOs and BLPs. The proposed algorithm of classification of observations is described according to the following steps:

Step I Identify suspected VOs denoted by *V* set by employing the robust Reweighted Least Squares (RLS) based on Least Median of Squares (LMS). *Step II* Detect the suspected HLPs denoted as *S* set by using the Diagnostic Robust Generalized Potential based on Index Set Equality (DRGP (ISE)). *Step III* Form a deletion group/set *D* based on the union of *V* set and *S* set and label the remaining data *as* R set. *Step IV* Following the idea of Rahmatullah Imon (2005), the FIMGT is defined as in (2):

$$\mathrm{FIMGT_i} = \begin{cases} \dfrac{\hat{\varepsilon}_{\mathrm{i.R}}}{\hat{\sigma}_{\mathrm{R-i}}\sqrt{1 - \mathrm{w}^*_{\mathrm{ii.R}}}} & \text{for } \mathrm{i} \in \mathrm{R} \\ \dfrac{\hat{\varepsilon}_{\mathrm{i.R}}}{\hat{\sigma}_{\mathrm{R}}\sqrt{1 + \mathrm{w}^*_{\mathrm{ii.R}}}} & \text{for } \mathrm{i} \notin \mathrm{R} \end{cases} \qquad (2)$$

where $(\hat{\beta}_R)$, the parameter estimates, residuals $(\hat{\varepsilon}_{i.R})$, hat values $(w^*_{ii.R})$, standard deviation $(\hat{\sigma}_R)$ and standard deviation with the ith case deleted $(\hat{\sigma}_{R-i})$ are computed using the OLS to the remaining data, i.e. R set.

Any observation which corresponds to FIMGT which is larger than its cutoff point (CP_{FIMGT}) is considered as vertical outlier. The cutoff point, is defined as follows:

$$\mathrm{CP_{FIMGT}} = \mathrm{Median(FIMGT_i)} + 3\ \mathrm{MAD(FIMGT_i)} \qquad (3)$$

The following guideline depicts the classification of observations into four categories taking the idea of Alguraibawi et al. (2015) and Bagheri and Midi (2016) with slide modifications:

An Observation is defined as Regular Observation (RO) if
$|FIMGT_i| \leq CP_{\mathrm{FIMGT}}$ and $p_{ii} \leq Median\ (p_{ii}) + cMAD\ (p_{ii})$

An Observation is defined as Vertical Outlier (VO) if
$|FIMGT_i| > CP_{\mathrm{FIMGT}}$ and $p_{ii} \leq Median\ (p_{ii}) + cMAD\ (p_{ii})$

An Observation is defined as a GLP if
$|FIMGT_i| \leq CP_{\mathrm{FIMGT}}$ and $p_{ii} > Median\ (p_{ii}) + cMAD\ (p_{ii})$

An Observation is declared as a BLP if
$|FIMGT_i| > CP_{\mathrm{FIMGT}}$ and $p_{ii} > Median\ (p_{ii}) + cMAD\ (p_{ii})$

where *c* is a selected constant such as 2 or 3. The resultant categories are presented in Figure 1 whereby observations are separated into regular observations (ROs), VOs, GLPs, and BLPs.

FIMGT \ DRGP		
	VOs	BLPs
	ROs	GLPs
	VOs	BLPs

FIGURE 1. Classification of observations into 4 categories

Hence, our proposed initial weight only considers those observations that fall into classes of VOs and BLPs but not in the class of ROs and GLPs. Thus, our propose initial weight is given by

$$d_i = \min\left[1, \left(\frac{\mathrm{CP}_{\mathrm{FIMGT}}}{FIMGT}\right)\right], i = 1,2,\ldots,n \qquad (4)$$

where $\mathrm{CP}_{\mathrm{FIMGT}}$ is defined as in Equation (3).

THE ALGORITHM OF INDEX SET EQUALITY

The Index Set Equality (ISE) which is another new technique from fast MCD (Salleh 2013) is used as an alternative to MVE or MCD. ISEs' running time is very fast because the algorithm of ISE only takes into account a comparison of two index set. Let $I_{old} = \{\pi_{(1)}^{old}, \pi_{(2)}^{old}, \ldots, \pi_{(h)}^{old}\}$ be the index set consisting of observations in a sample, denoted as H_{old}. Let $I_{new} = \{\pi_{(1)}^{new}, \pi_{(2)}^{new}, \ldots, \pi_{(h)}^{new}\}$ be the index set comprising of observations in another sample, denoted as H_{new}. The algorithm of ISE is given as follows:

Step 1 An arbitrary subset, H_{old} comprises of h different observations are chosen where h is smallest integer greater than or equal to $\frac{n+p+1}{2}$, p is the number of predictor variables (Rouseeuw & Driessen 1999). *Step 2* Compute the average vector $\bar{T}_{\mathrm{H_{old}}}$ and covariance matrix $\mathrm{C}_{\mathrm{Hold}}$ of all observations that belong to $\mathrm{H}_{\mathrm{old}}$. *Step 3* Compute the Mahalanobis Distance Squares, denoted as: $d_{old}^2(i) = (t_i - \overline{T}_{\mathrm{H_{old}}})' C_{\mathrm{H_{old}}}^{-1}(t_i - \overline{T}_{\mathrm{H_{old}}})$ for i = 1,2,...,n. *Step 4* Arrange $d_{old}^2(i)$ for i = 1,2,...,n in ascending order $d_{old}^2(\pi(1)) \leq d_{old}^2(\pi(2)) \leq \ldots \leq d_{old}^2(\pi(n))$ where π is permutation equal to {1,2,…,n}. *Step 5* Create $H_{new} = \{t_{\pi(1)}, t_{\pi(2)}, \ldots, t_{\pi(h)}\}$ such that its' elements comprises of the first smallest h observations acquired from Step 4. Then list the new Index Set, . *Step 6* Compare $I_{new} = I_{old}$. If $I_{new} = I_{old}$, stop the process. Afterwards, equate $\bar{T}_{H_{old}} := \bar{T}_{H_{new}}$, $C_{H_{old}} := C_{H_{new}}$, if $I_{new} \neq I_{old}$ then recompute $\bar{T}_{H_{new}}$, and $C_{H_{new}}$, let $H_{old} := H_{new}$,$\bar{T}_{H_{old}} := \bar{T}_{H_{new}}$ and $C_{H_{old}} := C_{H_{new}}$. Repeat Steps 3-6, until $I_{new} = I_{old}$ where at this point, $\bar{T}_{H_{new}}$ is the robust estimator of location and $C_{H_{new}}$ is the robust estimator of scatter.

THE PROPOSED GM-FIGMT ESTIMATOR

The algorithm of GM-FIGMT is summarized below:
Step 1 Calculate the residuals (r_i) based on S estimator developed by Rousseeuw (1984). *Step 2* Calculate the estimated scale (σ) of the residuals s = (1.4826)(*the median of the largest* (*n-p*) *of the* $|r_i|$), where r_i is obtained from Step 1.
Step 3 Compute the standardized residuals (e_i), where, $e_i = r_i/s$. *Step 4* Calculate the initial weight, denoted as d_i, where $d_i = min\,[1, \frac{\mathrm{CP_{FIMGT}}}{FIMGT}]$. *Step 5* Compute the bounded influence function for bad leverage points, $t_i = e_i/d_i$. *Step 6* Employ the weighted least squares (WLS) to estimate the parameters of the regression, $\hat{\beta} = (X^T W X)^{-1} X^T W Y$, where the weight w_i is reduced for large residuals to get good efficiency (In this paper, Tukey weight function is employed). *Step 7* Calculate the new residuals (r_i) from WLS and repeat steps (2-6) until convergence.

SIMULATION STUDY

A simulation study is conducted to investigate the performances of our proposed GM-FIMGT at various contamination scenarios. Linear regression model with three explanatory variables (p=3) are generated according to the following relation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + r_i$$

where r_i is the error term distributed as $N(0, 1)$, x_1, x_2 and x_3 are generated from $N(0,1)$. In this simulation study, we consider three contamination scenarios namely, BLPs, combination of GLPs and VOs, combinations of GLPs, BLPs and VOs. The contamination are created by randomly replaced some good observations in variables x_1 and with arbitrarily large number equal to 100. For each scenario, we consider five samples of sizes 30, 50, 100, 150, 200, and two percentage of contaminations (α = 0.05, 0.10). The four methods, namely the OLS, MM. GM6 and GM-FIMGT were then applied to the data. Some summary statistics over 1000 replicates were computed, such as the mean estimated values.

$$\widehat{\overline{B}}_j = \frac{1}{1000}\sum_{k=1}^{1000}\hat{\beta}_j^{(k)}$$

Following Riazoshams and Midi (2016), the performance of each technique is evaluated based on the percentage of robustness measures or efficiency using the ratio of the MSEs of the estimators compared with the MSEs of the OLS estimator of the good data. To simplify presentation, we report the efficiency based on overall MSE as follow:

$$MSE = \frac{1}{n}\left\|Y - X\hat{\beta}\right\|^2$$

The biases are also used as another criterion for evaluating the performances of the estimators. The overall bias is defined as

$$Bias = \sum_{j=0}^{p}(\hat{\beta}_j - \beta_j)^2$$

All computations were done using R Programming Language. The results are exhibited in Tables 1-3. The biases are shown in parenthesis. A good method is the one that has the highest value of efficiency and the least value of bias. Highest efficiency value implies that the MSE of the proposed method is closest to the MSE of the OLS for clean data, compared to other estimators considered in this study. Due to space limitations, we only present the results for p equals 3. However, other results were consistent.

TABLE 1. Overall Efficiency and biases for bad leverage points

N	OLS	MM	GM6	GM-FIMGT
		5% (BLP)		
30	22.1022 (2.2206)	90.0375 (0.0044)	**91.7436** **(0.0116)**	90.6562 (0.00311)
50	24.8685 (1.1018)	87.0163 (0.0120)	**94.5888** (0.0033)	93.7204 (0.0159)
100	16.92737 (1.0316)	90.1319 (0.0074)	**93.5777** **(0.0124)**	91.4400 (0.0068)
150	14.4951 (1.0267)	93.85369 (0.00573)	94.49078 (0.0135)	**95.1060** **(0.0080)**
200	12.2301 (1.1177)	94.7470 (0.0026)	94.0099 (0.0035)	**94.3328** **(0.0017)**
		10% (BLP)		
30	30.9373 (1.3677)	54.8025 (0.2555)	**87.4720** **(0.0201)**	87.3420 (0.1089)
50	25.4902 (1.092445)	65.0690 (0.07121)	**92.4413** **(0.0182)**	92.2498 (0.0214)
100	17.0289 (1.0297)	67.5062 (0.02778)	**91.1059** **(0.0170)**	89.7283 (0.0069)
150	10.2627 (2.10774)	89.8039 (0.0053)	91.8940 (0.0151)	**94.1150** **(0.0064)**
200	12.1874 (1.1071)	87.7782 (0.0070)	90.83337 (0.0117)	**91.8532** **(0.0011)**

TABLE 2. Overall Efficiency and biases for combination GLPs and VOs

N	OLS	MM	GM6	GM-FIMGT
		5% (GLP&VO)		
30	30.4603 (1.4095)	79.2962 (0.041908)	92.7848 (0.0128)	**97.461** **(0.0166)**
50	12.9200 (2.7561)	**103.6379** **(0.0156)**	89.5681 (0.1290)	102.1948 (0.0229)
100	4.8933 (3.7537)	100.1236 (0.0376)	86.6705 (0.0938)	**100.9936** **(0.0381)**
150	9.2348 (2.1467)	104.081 (0.0417)	92.3589 (0.0525)	**105.8981** **(0.0372)**
200	9.2512 (1.8624)	111.4084 (0.0180)	92.3589 (0.0525)	**113.9356** **(0.0235)**
		10% (GLP&VO)		
30	7.1508 (6.4014)	106.4538 (0.0134)	79.7749 (0.1996)	**107.4581** **(0.0101)**
50	19.7272 (1.8028)	100.8558 (0.0553)	91.6703 (0.0749)	**100.8836** **(0.0522)**
100	18.8543 (1.2766)	103.8755 (0.0794)	91.6330 (0.0562)	**103.8818** **(0.0770)**
150	7.4398 (2.6678)	103.3687 (0.0869)	91.8135 (0.0583)	**105.5938** **(0.0847)**
200	12.6561 (1.2346)	105.1071 (0.0471)	92.1644 (0.0408)	**109.1424** **(0.0490)**

TABLE 3. Overall Efficiency and biases for combination of GLPs, BLPs and VOs

N	OLS	MM	GM6	GM-FIMGT
		5% (GLP, BLP&VO)		
30	5.4755 (8.4502)	105.5742 (0.012418)	87.9806 (0.115971)	106.7969 (0.0088)
50	10.4628 (3.2763)	96.8607 (0.024607)	89.8900 (0.1299)	100.7797 (0.0274)
100	8.4375 (2.7972)	102.0166 (0.0242)	91.2501 (0.0576)	102.6391 (0.0178)
150	14.5939 (1.0895)	98.3847 (0.0557)	95.4568 (0.0268)	100.8252 (0.0482)
200	7.7701 (2.2262)	105.8062 (0.0282)	93.2330 (0.0282)	106.6786 (0.0286)
		10% (GLP, BLP&VO)		

30	23.3419 (1.8701)	119.0781 (0.0466)	84.0534 (0.1634)	120.6089 (0.0348)
50	7.5782 (5.5210)	88.6558 (0.2103)	81.5265 (0.0086)	94.6960 (0.0392)
100	16.0229 (1.2988)	103.8066 (0.0359)	92.3925 (0.0321)	107.0608 (0.0259)
150	13.6582 (1.2988)	102.5125 (0.0527)	94.1327 (0.0268)	102.9573 (0.0397)
200	11.8793 (1.1824)	110.6358 (0.0246)	92.3377 (0.0270)	112.2312 (0.0228)

Tables 1 to 3 present very interesting results. In all contamination scenarios, the GM-FIMGT consistently shows the best performance compared to other methods. It can be seen that the efficiencies and biases of the GM-FIMGT are consistently the highest and the smallest, respectively compared to other estimators. This is due to the fact that GM-FIMGT is based on FIMGT which successfully detect VOs and BLPs and subsequently they are down weighted. On the contrary, the GM6 which is based on RMD-MVE not only suffers from swamping effect, but also, only able to detect high leverage points which includes both GLPs and BLPs. Thus, some GLPs are given smaller weight. It is interesting to observe that when only BLPs are present in the data (Table 1), as expected the GM6 is fairly close to GM-FIMGT.

REAL EXAMPLE: AIR CRAFT DATA SET

The merit of the proposed GM-FIMGT estimator is illustrated using Air-Craft Data set which is taken from Gray (1985). This data set consists of 23 observations where cost is the response variable and four predictor variables namely aspect ratio, life to drag ratio, weight of the plane and maximal thrust. The evaluation is based on the standard deviation of the estimates (SE). Since the distribution of the GM-FIMGT is intractable, bootstrap method is used to find the standard deviation of its estimates. One thousands bootstrapped samples are utilised in this regards. Firstly, we want to apply the classification algorithm to the data.

The number of GLPs and BLPs detected by RMD (MVE) and FIMGT is presented in Figure 2. It can be seen from Figure 2 that the RMD (MVE) detects 2 GLPs (cases 14 and 20), one BLP (case 22) and one VOs (case 16). On the other hand, the FIMGT detects two observations (case 21 and case 19) as GLPs, one BLP (case 22) and one VO (case 16). The number of detected BLPs and VOs will be utilised to determine the initial weights for GM-FIMGT while the GM6s' initial weight only depends on the number of detected HLPs irrespective whether they are GLPs or BLPs. The parameter estimates and biases (in parenthesis) of the four methods are exhibited in Table 4. We anticipated that the results of GM6 will be affected since GM6 not only gives smaller weight to BLPs, but it also incorrectly gives smaller weight to GLPs as well. However, the GM-FIMGT correctly give smaller weight only to BLPs, VOs and weight equals one to the GLPs.

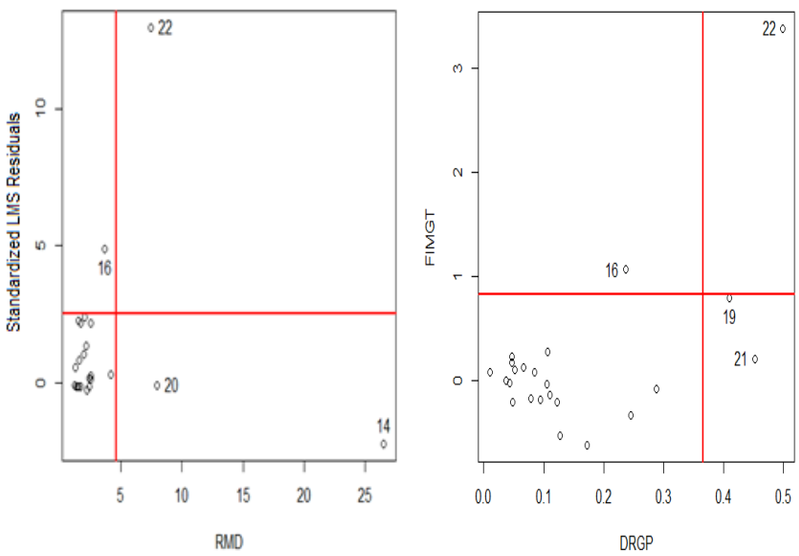

FIGURE 2. Classification of observations using RMD and FIMGT for aircraft data

TABLE 4. The parameter estimates, and bootstrap sd for aircraft data

Methods	Intercept	Aspect Ratio	Life to Drag Ratio	Weight Of Plane	Maximal Thrust
OLS	-3.79138 (8.61181)	-3.85292 (1.55736)	2.48827 (1.04388)	0.00350 (0.00042)	-0.00195 (0.00065)
M	6.14170 (5.70053)	-3.23057 (1.02219)	1.67112 (1.09407)	0.00192 (0.00030)	-0.00093 (0.00029)
GM6	9.92720 (6.67298)	-3.36519 (1.07516)	2.42709 (2.13377)	0.001432 (0.00044)	-0.00077 (0.00036)
GM-FIGMT	9.73986 (4.72351)	3.10215 (0.80041)	1.23737 (0.92506)	0.00140 (0.00033)	0.00058 (0.00028)

Table 4 clearly indicates that the OLS performs poorly. It can be observed that the OLS estimates have the largest standard errors. On the other hand, as can be expected, the GM-FIMGT is superior compared to GM6, MM and OLS estimators, evident by having the smallest standard error of the estimates. The results suggest that the GM-FIMGT did remarkably well when compared to other methods and it is consistent with the results of the simulation study.

Conclusion

The OLS is inefficient when outliers are present in a data. As an alternative, robust methods are put forward to remedy this problem. However, most robust methods such as the MM estimator have high breakdown point, high efficiency but are not robust against HLPs. The GM6 is the commonly used GM estimator which is robust against HLPs. Nonetheless its efficiency is affected when GLPs are present in a data set because the GM6 is based on RMD-MVE which is not capable of classifying observations into GLPs and BLPs. As such GLPs are given low weight.

We developed a new GM estimator in this regard to increase its efficiency. Our newly developed GM-FIMGT is based on FIMGT method which correctly identify VOs and BLPs. Hence, it is very successful in reducing only the effect of VOs and BLPs. The results of real data and simulation studies signify that our proposed GM-FIMGT method is more efficient than the existing methods in this study.

References

Alguraibawi, M., Midi, H. & Rahmatullah Imon, A.H.M. 2015. A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering* 2015: Article ID. 279472.

Andersen, R. 2008. *Modern Methods for Robust Regression - Series: Quantitatives Applications in Social Sciences.* United States of America: SAGE Publications, Inc. p. 152.

Bagheri, A. & Midi, H. 2016. Diagnostic plot for the identification of high leverage collinearity-influential observations. *SORT-Statistics and Operations Research Transactions* 39(1): 51-70.

Chatterjee, S. & Hadi, A.S. 2006. *Regression Analysis by Example*. 4th ed. Hoboken, New Jersey: John Wiley & Sons, Inc. pp. 21-45.

Coakley, C.W. & Hettmansperger, T.P. 1993. A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association* 88(423): 872-880.

Gray, J.B. 1985. Graphics for regression diagnostics. In *American Statistical Association Proceedings of the Statistical Computing Section Washington,* DC: American Statistical Association. pp. 102-107.

Hekimoğlu, S. & Erenoglu, R.C. 2013. A new GM-estimate with high breakdown point. *Acta Geodaetica et Geophysica* 48(4): 419-437.

Hill, R.W. & Paul, W.H. 1977. Two robust alternatives to least-squares regression. *Journal of the American Statistical Association* 72(360a): 828-833.

Huber, P.J. 2004. *Robust Statistics*. Hoboken, New Jersey: John Wiler & Sons, Inc. pp. 43-72.

Leroy, A.M. & Rousseeuw, J.P. 1987. *Robust Regression and Outlier Detection.* Hoboken, New Jersey: John Wiler & Sons, Inc. pp. 21-74.

Lim, H.A. & Midi, H. 2016. Diagnostic robust generalized potential based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. *Computational Statistics* 31(3): 859-877.

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

Rahmatullah Imon, A.H.M. 2005. Identifying multiple influential observations in linear regression. *Journal of Applied Statistics* 32(9): 929-946.

Riazoshams, H. & Midi, H. 2016. The performance of a robust multistage estimator in nonlinear regression with heteroscedastic errors. *Communications in Statistics-Simulation and Computation* 45(9): 3394-3415.

Rousseeuw, P.J. 1985. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications* 8(37): 283-297.

Rousseeuw, P.J. 1984. Least median of squares regression. *Journal of the American Statistical Association* 79(388): 871-880.

Rousseeuw, P.J. & Croux, C. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88(424): 1273-1283.

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

Salleh, R. 2013. A robust estimation method of location and scale with application in monitoring process variability. Universiti Teknologi Malaysia. Ph.D. Thesis (Unpublished).

Simpson, D.G., Ruppert, D. & Carroll, R.J. 1992. On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association* 87(418): 439-450.

Stromberg, A.J., Hössjer, O. & Hawkins, D.M. 2000. The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association* 95(451): 853-864.

Wilcox, R.R. 2005. *Introduction to Robust Estimation and Hypothesis Testing.* 2nd ed. Burlington, USA: Elsivier Inc. pp. 413-464.

Yohai, V.J. 1987. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics* 15(2): 642-656.

Yohai, V.J. & Zamar, R.H. 1988. High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* 83(402): 406-413.

Habshah Midi* & Jayanthi Arasan
Faculty of Science and Institute for Mathematical Research
Universiti Putra Malaysia
43400 UPM Serdang, Selangor Darul Ehsan
Malaysia

Shelan Saied Ismaeel
Department of Mathematics
Faculty of Science
University of Zakho
Iraq

Mohammed A Mohammed
Al-Dewanyia Technical Institute, AUT
Iraq

*Corresponding author; email: habshah@upm.edu.my

Received: 1 April 2020
Accepted: 9 August 2020