Journal of Mathematics and Statistics 5 (4): 311-321, 2009
ISSN 1549-3644

Robust Estimations as a Remedy for Multicollinearity Caused by Multiple High Leverage Points

Arezoo Bagheri and Habshah Midi
Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Abstract. **Problem statement:** The Least Squares (LS) method has been the most popular technique for estimating the parameters of a model due to its optimal properties and ease of computation. LS estimated regression may be seriously disturbed by multicollinearity which is a near linear dependency between two or more explanatory variables in the regression models. Even though LS estimates are unbiased in the presence of multicollinearity, they will be imprecise with inflated standard errors of the estimated regression coefficients. It is now evident that the multiple high leverage points which are the outliers in the X-direction may be the prime source of collinearity-influential observations. **Approach:** In this study, we had proposed robust procedures for the estimation of regression parameters in the presence of multiple high leverage points which cause multicollinearity problems. This procedure utilized mainly a one step reweighted least square where the initial weight functions were determined by the Diagnostic-Robust Generalized Potentials (DRGP). Here, we had incorporated the DRGP with different types of robust methods to downweight the multiple high leverage points which lead to reducing the effects of multicollinearity. The new proposed methods were called GM-DRGP-L_1, GM-DRGP-LTS, M-DRGP, MM-DRGP, DRGP-MM. Some indicators had been defined to obtain the best performance robust method among the existing and new introduced methods. **Results:** The empirical study indicated that the DRGP-MM emerge to be more efficient and more reliable than other methods, followed by the GM-DRGP-LTS as they were able to reduce the most effect of multicollinearity. The results seemed to suggest that the DRGP-MM and the GM-DRGP-LTS offers a substantial improvement over other methods for correcting the problems of high leverage points enhancing multicollinearity. **Conclusion/Recommendations:** In order to solve the multicollinearity problems which are mainly due to the multiple high leverage points, two proposed robust methods, DRGP- MM and the GM-DRGP-LTS, were recommended.

Key words: Multicollinearity, Multiple high leverage points, robust estimations, diagnostic robust generalized potentials method

INTRODUCTION

Least squares estimation is one of the predominant regression analysis techniques due to the universal acceptance, elegant statistical properties and computational simplicity. Unfortunately, the mathematical elegance that makes least squares so popular depends on a number of fairly restrictive and often unrealistic assumptions. Two of the assumptions that make least squares so attractive in terms of general model hypothesis and parameter significance testing, are normality of error distribution and independency of explanatory variables. The normality assumption can be violated in the presence of one or more sufficiently outlying observations in the data set resulting in less reliable estimates of the model parameters[1]. The second condition that potentially impacts the reliability of least squares estimation is multicollinearity, which is a near-linear dependency among the explanatory variables (X-direction). Multicollinearity can cause large variability in the estimation of parameters. Sometimes it causes the parameters estimation to be different from the true values by orders of magnitude or incorrect sign. It may also inflate the variance of the estimations. High leverage points, the points far from the rest of the data in the X-direction, have high potential for influencing most of the regression results such as eigenstructure and condition index of X[10].

Corresponding Author: Arezoo Bagheri, Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia
Tel: 603-8946-6876 Fax: 603-8942-3789

Hadi[10] noted that collinearity-influential points are usually the points with high-leverage which tends to pull the model fit to their direction. Kamruzzaman and Imon[17] introduced these points as a new source of multicollinearity problems. Thus, diagnosing the multiple high leverage points and recognizing estimation methods which are resistant to these points may improve regression estimations [28]. Robust regression methods are designed to be less sensitive than least squares to outliers mostly in Y-direction, resulting in improved fits to the non-outlying observations. In order to achieve this stability, robust regression limits the influence of outliers. Three most important properties of any robust regression are efficiency, breakdown point and bounded influence[21]. Several works on robust estimation have been proposed in the literature[2,3,12]. Among them are Huber[15] and Yohai[29] who introduced the M- and MM- estimators. However, the M-estimator is not robust in the X-direction and has a low break down point that is equal to (1/n)[27]. The MM- estimator has high efficiency and also possesses high breakdown values. Rousseeuw [23,24] introduced the Least Median of Squares (LMS) and Least Trimmed of Squares (LTS) in which both estimators have high breakdown equal to 50%. However, they are unbounded influence estimators [23,24] where the LMS and LTS have low and medium efficiency value, respectively[26]. Rousseuw and Leroy[25] proposed Reweighted Least Squares based on LMS(RLS-LMS) where the LMS scale employed to standardized the LS residuals and a hard rejection function utilized to assigned the initial weights to the data. Schweppe[13] introduced a class of robust methods which is called the Generalized M-estimators (GM-estimators) with a major aim of downweighting those high leverage points which have large residuals. Simpson[26] has reported that these estimators have high efficiency and bounded influence properties which achieve a moderate break down point equal to 1/p. The GM-estimator is the solution of the normal equation:

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{y_i - x_i'\hat{\beta}}{s\pi_i}\right) x_i = 0 \quad (1)$$

Where:

π_i = Defined to downweight high leverage points with high residuals

s = A robust scale estimate

Iteratively Reweighted Least Squares (IRLS) may be used to solve (1). At convergence, the GM-estimator may be written:

$$\hat{\beta}_{GM} = (X'WX)^{-1}X'Wy \quad (2)$$

where, in this case the diagonal elements of W are the weights w_i defined as:

$$w_i = \frac{\psi\left[(y_i - x_i'\hat{\beta}_{GM})/\pi_i s\right]}{(y_i - x_i'\hat{\beta}_{GM})/\pi_i s} \quad (3)$$

The main objective of this study is to propose some estimators that are able to perform well where multiple high leverage points are the cause of the multicollinearity problems in regression analysis. Nonetheless, the development of such estimators has not been published extensively in the literature. Since high leverage points may be collinearity-enhancing observations, we attempt to reduce its influence by employing robust estimator which is known to be resistant to high leverage points. In this connection, we will consider the bounded influence or Generalized M-estimators[13] with a major aim of down weighting those high leverage points which have large residuals. To enhance the GM-estimators, these estimators may be defined as multi-stage estimators where in different stages, different robust techniques are applied to combine the desirable properties of each technique[7,27]. Hence in this study, we propose mainly new multi stage GM-estimators and weighted MM-estimators to remedy the problem of collinearity-enhancing observations on the parameter estimates of the multiple linear regression model.

MATERIALS AND METHODS

High leverage points diagnostic methods: A traditional measure of the outlyingness of an observation X_i with respect to the sample is three-Sigma edit rule which is defined as follows:

$$T = \frac{X - \hat{X}}{s} \quad (4)$$

Where:

$\hat{X}$ = The mean of explanatory variables

s = The standard deviation of explanatory variables

The robust version of (4) is:

$$T' = \frac{X - Med(X)}{Mad(X)} \quad (5)$$

Where:

Med(X) = Median(X)

Mad(X) = The normalized median absolute deviation about the Median(X) (Mad = 1.4826(median $|x_i$-median $(x_i)|$)

When the distribution of the data is normal, T and T′ are approximately equal. Any observation which has absolute value of , T or T' greater than 3, is considered as outlier[22]. This method can be used in univariate regression models as a diagnostics rule to detect high leverage points. Kamaruzzaman and Imon[17] pointed out that high leverage points are a new prime source of multicollinearity. It is now evident that high leverage collinearity-enhancing observations are those points in which their values are in large magnitude at least for two explanatory variables. Since in most of the regression analysis, more than one explanatory variable exists in the model, investigating some useful methods in these cases seems to be necessary. One of the handiest methods can be defined as hat matrix.

Hat matrix which is traditionally used as a measure of leverage points in regression analysis is defined as W =X $(X^T X)^{-1}X^T$. The most widely used cutoff point of the hat matrix is twice-the-mean-rule $(2k/n)$[14]. Nevertheless, Hadi[11] pointed out that the hat matrix may fail to identify the high leverage points due to the effect of high leverage points in leverage structure. So, he introduced another diagnostic tool as follows:

$$p_{ii} = \frac{w_{ii}}{1 - w_{ii}} \quad (6)$$

where, $w_{ii} = x_i^T (X^T X)^{-1} x_i$ is the diagonal element of W and the i-th, diagonal potential p_{ii} can be defined as: $p_{ii} = x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i$ where $X_{(i)}$ is the data matrix X without the i-th row. The proposed cut off point for potential values p_{ii} can be defined as Median (p_{ii})+c Mad (p_{ii}) (MAD-cutoff point) where c can be the constant values of 2 or 3. Still, this method was unable to detect all of the high leverage points.

Another diagnostic tool which is called generalized potential introduced by [16].Generalized potentials for the whole data set can be defined as:

$$p_{ii}^* = \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \text{ for } i \in R = w_{ii}^{(-D)} \quad \text{for } i \in D \quad (7)$$

Where:

D = The deleted set which corresponds to the suspected outliers

R = The remaining set from observations after deleting d < (n-k) which contains (n-d) cases

Since there isn't any finite upper bound for p_{ii}^*'s and the theoretical distribution of them are not easy to derived, he introduced a MAD-cutoff point for the generalized potential as well.

Recently, Habshah *et al.*[9] developed Diagnostic Robust Generalized Potential (DRGP) to determine outlying points in multivariate data set by utilizing the Robust Mahalanobis Distance (RMD) based on Minimum Volume Estimator (MVE). We refer this method as the DRGP (MVE). The generalized potentials in (7) are computed based on the set R (remaining set) and the set D (deletion set) obtained from the RMD-MVE. Here the RMD-MVE is used to identify the suspected high leverage points (set D) and then diagnostic approach is used to confirm our suspicion. We then used the MAD-cutoff point to see whether all members of the deletion set have potentially high leverage or not.

Rousseeuw[24] defined RMD-MVE as follows:

$$RMD_i = \sqrt{(X - T_R(X))' C_R(X)^{-1} (X - T_R(X))} \text{ for } i = 1, ..., n \quad (8)$$

where, $T_R(X)$ and $C_R(X)$ are robust locations and shape estimates of the MVE. The merit of DRGP (MVE) method is in the swamping of less good leverage as high leverage points as compared to the RMD-MVE.

Multi-stage GM-estimator: The Multi-Stage GM-estimator was developed to overcome the problem of low break down point of the GM-estimators [27]. These estimators may have high break down point if appropriate initial estimators are used. A good starting value is always important in an iterative scheme. Table 1 includes some of the existing Multi-stage GM-estimators. Walker [28] defined the π-weight function which is a typical of least squares outlier diagnostic DFFITS where it uses the least squares and a non-iterated MAD as scale estimator in the initial stage. The final estimate is obtained using fully iterated reweighted least squares. The first GM-estimator with high efficiency, high breakdown and bounded influence was proposed by[7]. To overcome the limitation of Walker's method in using LS as initial estimator, Coakley and Hettmansperger[7] proposed employing high breakdown LTS as initial estimator; LMS scale as scale estimate and the robust distance based on MVE as leverage estimates. The π-weight estimator is a ratio of the x^2 cutoff value to the squared Robust Distance. A one-step Newton Raphson is used as a convergence approach. Simpson[26] has investigated several types of Multi-stage GM-estimators by considering various outlier magnitudes through simulation studies. He verified that the combination of GM-and MM-estimators outperform other existing methods. The research in this area is at the early stage of emergence and some combinations of efficient π-weight and ψ-functions may produce excellent estimators (Table 1).

Table1: Some definition of existing GM-estimators

	Technique	
Component	Walker[28]	Coakley and Hettmansperger[7]
Initial estimate	LS	LTS
Scale estimate	MAD	LMS scale = 1.4826 $\left(1+\frac{5}{n-p-1}\text{Median}\left\|r_i\right\|\right)$
Leverage measure	h_{ii}	Robust mahalanobis distance (based on MVE)
π-weight function	$[(1-h_{ii})/(h_{ii})^{\frac{1}{2}}]$	$\min\left[1,\left\{\frac{X^2_{0.025,p-1}}{RD^2}\right\}\right]$
ψ-function	Huber	Huber
Tuning constant	1.345	1.345
Convergence approach	Fully iterated IRLS	One-step Newton Raphson

Proposed Multi-stage estimator: One of the drawbacks of the existing GM-estimator is in the definition of π-weight that depend on Robust Distance based on MVE which tends to swamp some low leverage points even though it can identify high leverage points correctly. Thus, it will produce low weights to some of the good leverages as well [9]. In this connection the precision of the GM-estimator can be improved by utilizing more effective diagnostics method. Subsequently an efficient π-weight function is obtained. This motivates us to consider the DRGP which is proposed by Habshah *et al*.[9] in the calculation of the π_i-function. The attractive feature of the DRGP is that it is very successful in identifying multiple high leverage points and swamps less good leverages as high leverage points when compared to RMD-MVE. In this study, we proposed Multi-stage GM-estimators by incorporating the GM-estimators with slight modification in which the DRGP proposed by Habshah *et al*.[9] is employed in the computation of the π_i. Here, the GM- and MM-estimators are considered because Simpson *et al*.[27] enumerated that these estimators surpass other robust methods. The first two proposed estimators are Multi-stage GM-estimators while the others are defined based on the M- and MM-estimator.

It is important to point out that in the new proposed methods, the DRGP statistics is referred as P_i with MAD-cutoff points. Here, we will employ the Tukey's biweight redescending ψ-function[4] which is defined as:

$$\psi(t)=\begin{cases} t\left[1-\left(\frac{t}{c}\right)^2\right]^2 & \text{if } |t|\le c \\ 0 & \text{if } |t|>c \end{cases} \qquad (9)$$

The Tukey's biweight with the tuning constant c = 4.685 will result a 95% efficiency under normal error distribution. A redescending ψ-function is better comparing to monotonic functions such as Huber's function because the former assigns lower weights (even zero if the residual is too large) to large outliers. In this respect, a redescending ψ-functions limits the influence of outliers more effectively than a monotone ψ-function. The proposed methods can be computed in three steps and summarized as follow.

GM-DRGP-L_1:

Step 1: Employ L_1 estimators as initial estimate and then obtain the standardized residuals of L_1 estimator. Compute MAD = 1.4826 (med$|r_i$-med$(r_i)|$) for non-zero residuals according to Maronna and Yohai[20]. It is important to mention that if MAD is computed from all the residuals of L_1 estimators, the scale estimates will become too small due to defining some zero residual. Thus, non-null residuals have been used to compute the scale estimate.

Step 2: Define $\pi_i=\min\left[1,\frac{\text{MAD}-\text{cutoff}(p_i)}{(p_i)}\right]$ in (1) and use function (9) to assign final weights to the observations.

Step 3: Compute a one step reweighted least squares as a convergence approach.

GM-DRGP-LTS:

Step 1: Consider the LTS as initial estimate and compute the standardized residuals and scale estimate based on LTS.

Step 2: Define $\pi_i=\min\left[1,\frac{\text{MAD}-\text{cutoff}(p_i)}{(p_i)}\right]$ in (1) and use function (9) to assign final weights to the observations.

Step 3: Compute a one step reweighted least squares as convergence approach.

M-DRGP:
Step 1: Compute the residuals of M-estimators scale by assigning the initial weight of W_i (DRGP(MVE)) = $\min\left[1, \frac{MAD - cutoff(p_i)}{(p_i)}\right]$ where P_i is DRGP(MVE) statistics.

Step 2: Define new weights as $w_i = r_i$ (M-estimator)/scale (M-estimator) and use a Tukey's biweight to assign final weight to the observations.

Step 3: Compute a one step reweighted least squares.

MM-DRGP: This method is similar to that of M-DRGP, where on the second and third steps the M-estimator is replaced with the MM-estimator.

DRGP-MM:
Step 1: Compute the initial weight W_i (DRGP(MVE)) which is defined in the first step of M-DRGP and using function (9) to assign final weights to the observations.

Step 2: Compute the weighted MM-estimators by these final weights.

Weighted multicollinearity diagnostics: Weighted multicollinearity diagnostics are defined as practical tools to investigate the source of multicollinearity which may be the high leverage points in the data set. Indeed, robust estimators to deal with multicolinearity problems are largely ignored issues. Walker[28] noted that sometimes the weighting process in robust methods can decrease the multicollinearity of X matrix. An effective measure of robust methods which reduce multicollinearity problems due to the presence of multiple high leverage points can be defined as weighted multicollinearity diagnostics. The two most classical and practical multicollinearity diagnostics are Correlation X matrix and Variance Inflation Factors (VIF)[6]. In bivariate regression analysis, when correlation coefficient exceeds 0.9 multicollinearity can be detected. However, in the case of more than two explanatory variables model, multicollinearity may occur in less than 0.9 correlation coefficients [22]. Since, this multicollinearity diagnostics is simple and easy to compute, it is more preferred [6]. Another practical approach to detect multicollinearity is by using Variance Inflation Factors (VIF). VIF is defined as $VIF(i) = (1-R_i^2)^{-1}$ where R_i is the coefficient determination of regressing each X_i on the other explanatory variables, which produced a valuable indices to detect inflated variances of regression parameter estimations[19]. Cutoff point of 5 and 10 are recommended as a rule of thumb for VIF to detect moderate and severe multicollinearity, respectively. The weighted linear regression can be expressed as a transformed model [18]:

$$Y_w = X_W \beta + \varepsilon_w \quad (10)$$

Where:
$Y_w = W^{1/2}Y$, $X_w = W^{1/2}X$ and $\varepsilon_w = W^{1/2}\varepsilon$

The final weights of the proposed estimators, which are expected to be robust against high leverage points, can be used in the computation of weighted multicollinearity diagnostics. These diagnostics can be defined as a measure to evaluate which method is more robust against the high leverage points which are responsible for the multicollinearity. It is important to note that all high leverage points are not collinearity-influential and vice versa[11]. The weighted correlation matrix can be computed through the correlation matrix of X_w. The weighted VIF is defined as follows:

$$VIF_w(i) = (1-R_i^2(w))^{-1} \quad (11)$$

where, $R^2(W)$ is the coefficient of determination of regressing each X_{wi} on the other weighted explanatory variables. It is worth mentioning that if the high leverage points are the source of multicollinearity in the data set, the weighted multicollinearity diagnostics will not detect multicollinearity due to these points otherwise multicolliearity will be detected easily.

RESULTS

Numerical example: To evaluate the performance of our proposed robust methods a real data set is considered.

Child mortality data set: Gujarati[8] introduced this data set with 64 observations which includes child mortality as dependent variable and Gross National Production (GNP) per capita and Female Literacy Rate (FLR) as independent variables. Table 2 presents the classical multicollinearity diagnostics methods such as the correlation matrix and VIF. It is important to note here that the multiple high leverage points will be the prime source of multicollinearity when they are detected as high leverage points with the large magnitude in at least two explanatory variables. The diagnostic methods of hat matrix, DRGP(MVE) and robust three-sigma edit rule (Eq. 5) for the original and modified data set are shown in Table 3.

Table 2: Multicollinearity diagnostics and least square coefficients of original and modified child mortality data set

	$Cor(X_1,X_2)$	VIF	b_1(t p-value)	b_2 (t p-value)	F p-value	S(e)
Original data	0.27	1.080 (0.007)	-2.230 (0.000)	-0.010	0.000	41.75
Modified data	0.99	37.340	-0.240 (0.512)	0.002 (0.938)	0.003	70.25

Table 3: High leverage diagnostics for original and modified child mortality data set

Original data

Index	hat(X)(0.09)	DRGP(X)(0.11)	$T'_1(3)$	$T'_2(3)$
1	0.02	0.20	0.34	2.16
5	0.03	0.14	0.89	2.47
24	0.05	0.90	1.03	6.26
27	0.05	0.35	1.22	4.15
30	0.77	31.67	0.47	33.22
33	0.14	5.52	0.06	13.87
38	0.05	0.15	1.25	2.54
53	0.05	1.02	0.97	6.59
54	0.05	0.14	1.10	0.60
58	0.07	0.91	1.47	6.48
62	0.05	0.59	1.16	5.21

Modified data

Index	Modified X_1	hat(X)(0.09)	DRGP(X)(0.11)	$T'_1(3)$	$T'_2(3)$
24	248	0.03	1.23	5.28	6.26
27	180	0.02	0.55	3.50	4.15
30	1107	0.75	34.72	28.01	33.22
33	490	0.13	6.04	11.69	13.87
53	258	0.04	1.36	5.56	6.59
58	255	0.11	0.14	5.47	6.48
62	214	0.03	0.85	4.39	5.21

Table 4: Multicollinearity diagnostics and least square coefficients of different methods for modified child mortality data set

Method	$Cor(x_1,x_2)>0.9$	VIF>5	b_1(t p-value)	b_2(t p-value)	F p-value	S(e)
Ls	0.99	37.34	-0.240 (0.512)	0.002 (0.938)	0	70.25
GM-DRGP-L_1	0.98	33.36	-0.680 (0.020)	-0.020 (0.340)	0	55.00
GM-DRGP- LTS	0.65	1.75	-1.780 (0.000)	-0.040 (0.000)	0	39.20
DRGP- MM	0.65	1.74	-1.610 (0.000)	-0.040 (0.000)	-	36.17
MM-DRGP	0.90	5.25	-0.690 (0.020)	-0.010 (0.330)	0	54.80
M-DRGP	0.90	5.25	-0.690 (0.020)	-0.010 (0.330)	0	54.80
RLS-LMS	0.65	1.74	-1.800 (0.000)	-0.040 (0.000)	0	39.45

In order to obtain a large magnitude of high leverage points in X_1 as in X_2, the value of T'_2 for X_2, should be equal to the value of T'_1 for X_1 as in Eq. 5. Since the observations 24, 27,30, 33, 53, 58 and 62 of variable X_2 were identified as high leverage points by both T'_2 and DRGP(MVE) methods, the variable X_1 for those observations should then be modified such that their values become in the same magnitude for both explanatory variables. The modified x_1 are as follows:

$$Modified(X_1) = T'_2*(Mad(X_1))+Median(X_1)$$

With these modifications, the high leverage points are referred as collinearity-enhancing observations. Table 4 presents the multicollinearity diagnostics and least squares coefficients of the modified child mortality data set for the proposed robust methods and the existing robust methods (Table 2-4).

Monte Carlo simulation: A simulation study has been carried out to further assess the performance of the proposed estimators in higher dimensions. In this study, we consider multiple linear regression model with moderate sample size equals to 100 and different number of explanatory variables, that is p = 3, 5 and 7. We set the p+1 true regression coefficients equal to one and consider a regression model with intercept. Following Rousseeuw and Leroy[25] simulation design, each of the p explanatory variables were generated from the multivariate normal distribution N (0, 100). Then we start to contaminate the data by generating leverage

points from N (100,100). For each variable, we generate high leverage point by deleting 'good' observation and replacing it with a high leverage point. The level of high leverage points varied from 0-50%. Here, we consider different error terms which are generated from standardized normal distribution, exponential distribution with mean equal to one and student distribution with 3 and 8 degree of freedoms. Belsley *et al*.[5] pointed out that the estimation of regression parameters is unbiased in the presence of multicollinearity problems. However, when the source of multicollinearity is multiple high leverage points the estimators will be bias. Simpson [26] recommended using several indicators of Average Mean Square Error of Estimation (AMSEE) performance to determine the best overall performance of the existing and proposed methods. The AMSEE is defined by:

$$\text{ASMEE} = \text{mean}\left[(\hat{\beta}_R - \beta)'(\hat{\beta}_R - \beta)\right] = \text{mean}(\text{MSEE}) \quad (12)$$

Where:

$\hat{\beta}_R$ = The parameter estimation of the proposed methods

β = The true parameter value

Thus after finding the MSEE for each generated sample size, we take the average in 1000 replications. Table 5 shows the AMSEE for a sample of size 100 with three explanatory variables and standardized normal distribution of error for different percentage of multiple high leverage points. Simpson[26] introduced three indicators to evaluate the AMSEE performance. The first indicator of AMSEE performance is the relative AMSEE rank of each technique on different levels of high leverage points and different error distributions. The AMSEE values are ranked from lowest to highest and the Summed Ranks (SR) of each technique are obtained. A lower rank indicates a better AMSEE performance. The second indicator of performance is the standard deviation of the ranks (RSD). This indicator is the most important criteria in assigning the final ranks to the similar ranks. Table 5 includes the AMSEE rankings of the data according to these two indicators (Rank (1)). However, standard deviation of the ranks in the final ranks (1) computation has not been utilized due to non-equality of the overall ranks in Table 5.

Table 5: AMSEE of n=100 and p=3 for standardized normal error terms

		AMSEE, n = 100, p = 3								
Distribution	HL (%)	Ls	M	M-DRGP	MM	MM-DRGP	GM-DRGP-L_1	GM-DRGP -LTS	RLS-LMS	DRGP-MM
	0	0.04	0.06	0.04	0.05	0.04	0.04	0.04	0.04	0.06
	10	3.01	0.06	0.05	0.06	0.05	3.06	0.05	0.05	0.06
	20	3.04	0.08	0.37	0.07	0.07	3.05	0.05	0.21	0.07
N(0,1)	30	3.05	0.76	2.79	0.75	2.60	3.06	0.06	0.26	0.08
	40	3.06	3.03	3.33	3.02	3.32	3.07	0.08	0.49	0.08
	50	3.04	3.06	3.51	3.05	3.49	3.05	0.22	0.66	0.09
		Performance rank AMSEE according to the first and second indicator n = 100, p = 3								
	0	1.00	3.00	1.00	2.00	1.00	1.00	1.00	1.00	3.00
	10	3.00	2.00	1.00	2.00	1.00	4.00	1.00	1.00	2.00
	20	6.00	3.00	5.00	2.00	2.00	7.00	1.00	4.00	2.00
N(0,1)	30	8.00	5.00	7.00	4.00	6.00	9.00	1.00	3.00	2.00
	40	5.00	4.00	8.00	3.00	7.00	6.00	1.00	2.00	1.00
	50	4.00	6.00	8.00	5.00	7.00	5.00	2.00	3.00	1.00
	SR	27.00	23.00	30.00	18.00	24.00	32.00	7.00	14.00	11.00
	Overall rank	7.00	5.00	8.00	4.00	6.00	9.00	1.00	3.00	2.00
	RSD	2.43	1.47	3.29	1.26	2.97	2.73	0.41	1.21	0.75
	Rank (1)	7.00	5.00	8.00	4.00	6.00	9.00	1.00	3.00	2.00
		Performance rank AMSEE according to the third indicator n = 100, p = 3								
	0	0	40	0	25	0	0	0	0	50
	10	7425	50	25	50	25	7550	25	25	50
	20	7500	100	825	75	75	7525	25	425	75
N(0,1)	30	7525	1800	6875	1775	6400	7550	50	550	100
	40	7550	7475	8225	7450	8200	7575	100	1100	100
	50	7500	7550	8675	7525	8625	7525	450	1550	125
	SUM	37500	17015	24625	16900	23325	37725	650	3650	500
	Rank (2)	8	5	7	4	6	9	2	3	1
Sum of rank	(1) and (2)	15	10	15	8	12	18	3	6	3
N(0,1)	Final ranks	6	4	6	3	5	7	1	2	1

HL (%): Percentage of high leverage points; SR: Sum of Ranks; RSD: Rank SD

Table 6: Performance rank (AMSEE) of different methods for n = 100 in different distribution of error terms and different number of explanatory variables

Distribution	Ls	M	M-DRGP	MM	MM-DRGP	GM-DRGP-L_1	GM-DRGP-LTS	RLS-LMS	DRGP-MM
	Performance rank (AMSEE) n = 100, p = 3								
N(0,1)	6.00	4.00	6.00	3.00	5.00	7.00	1.00	2.00	1.00
Exp(1)	8.00	4.00	5.00	2.00	7.00	9.00	3.00	6.00	1.00
t(1)	8.00	4.00	7.00	3.00	5.00	9.00	2.00	6.00	1.00
t(8)	6.00	3.00	5.00	2.00	4.00	6.00	1.00	5.00	1.00
Sum of rank	28.00	15.00	23.00	10.00	21.00	31.00	7.00	19.00	4.00
Overall rank	8.00	4.00	7.00	3.00	6.00	9.00	2.00	5.00	1.00
Rank SD	1.15	0.50	0.96	0.58	1.26	1.50	0.96	1.89	0.00
	Performance rank (AMSEE) n = 100, p = 5								
N(0,1)	6.00	5.00	6.00	4.00	6.00	7.00	1.00	3.00	2.00
Exp(1)	8.00	6.00	7.00	3.00	7.00	9.00	4.00	5.00	1.00
t(3)	6.00	5.00	6.00	4.00	6.00	7.00	2.00	4.00	1.00
t(8)	7.00	4.00	6.00	5.00	6.00	8.00	1.00	3.00	2.00
Sum of ranks	27.00	20.00	25.00	16.00	25.00	31.00	8.00	15.00	6.00
Overall rank	7.00	5.00	6.00	4.00	6.00	8.00	2.00	3.00	1.00
Rank SD	0.96	0.82	0.50	0.82	0.58	0.96	1.41	0.96	0.58
	Performance rank (AMSEE) n = 100, p = 7								
N(0,1)	6.00	5.00	6.00	4.00	6.00	7.00	2.00	3.00	1.00
Exp(1)	8.00	5.00	7.00	3.00	6.00	9.00	2.00	4.00	1.00
t(1)	6.00	4.00	5.00	3.00	5.00	7.00	2.00	4.00	1.00
t(8)	7.00	4.00	6.00	2.00	5.00	8.00	1.00	3.00	1.00
Sum of ranks	27.00	18.00	24.00	12.00	22.00	31.00	7.00	14.00	4.00
Overall rank	8.00	5.00	7.00	3.00	6.00	9.00	2.00	4.00	1.00
Rank SD	0.96	0.58	0.82	0.82	0.58	0.96	0.50	0.58	0.00

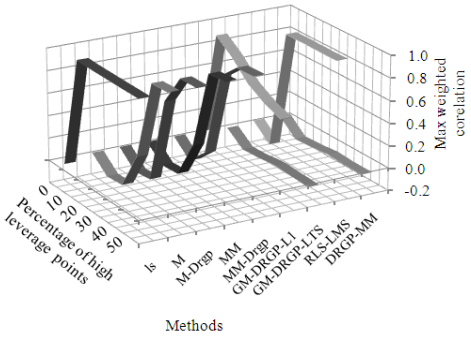
Table 7: Performance final ranks (AMSEE) of different methods for n = 100 in different number of explanatory variables

Distribution	Ls	M	M-DRGP	MM	MM-DRGP	GM-DRGP-L_1	GM-DRGP-LTS	RLS-LMS	DRGP-MM
Final rank p = 3	8.00	4.00	7.00	3.00	6	9.00	2	5.00	1
Final rank p = 5	7.00	5.00	6.00	4.00	6	8.00	2	3.00	1
Final rank p = 7	8.00	5.00	7.00	3.00	6	9.00	2	4.00	1
Sum of ranks	23.00	14.00	20.00	10.00	18	26.00	6	12.00	3
Final rank (3)	8.00	5.00	7.00	3.00	6	9.00	2	4.00	1
Rank SD	0.58	0.58	0.58	0.58	0	0.58	0	1.00	0
Final rank	8.00	5.00	7.00	3.00	6	9.00	2	4.00	1

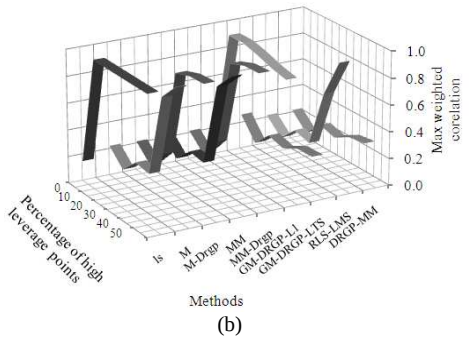
A third indicator of AMSEE performance involves accounting for the differing AMSEE ranges among techniques within different level of high leverage points and error distributions. According to Table 5, the spread between the highest and lowest AMSEE for 10% high leverage points and without high leverage points is 3.01 and 0.02 respectively. Thus, being ranked last in the situation without high leverage points may not be as harmful as being ranked last in 10% high leverage points. One way for capturing this spread within different level of high leverage points is to compute, for each level of high leverage points and each error distribution, the percent above the minimum AMSEE. Thus, according to Table 5, the smallest AMSEE is equal to 0.04 and the percent above the minimum AMSEE for a technique with AMSEE of 0.05 is 25%. The sum ranks of the percent above the minimum AMSEE of each technique are recorded which represents the third indicator of AMSEE performance (Rank (2) in Table 5). Therefore, Ranks in different levels of high leverage for each error distribution and each specific number of explanatory variables allocated by ranking rank (1) and (2) which are illustrated in Table 5.

The rank in different error distributions for specific number of explanatory variables can be computed in three steps:

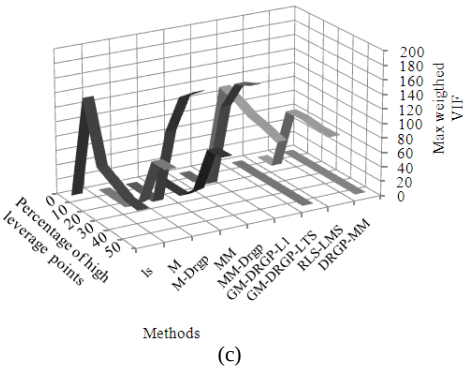
Step 1: rank the sum of the ranks for each error distribution in different level of high leverage points and for different number of explanatory variables by the three introduced indicators. It should be noticed that when the sum of the ranks is equal, the ranks assigned according to the standard deviations of the Ranks should be considered. For instance Table 5 presents the ranks of different estimators in different level of high leverage points when the error distribution is normal. The same procedure can be applied for error distributions of exponential with mean equal to one and t-student with 3 and 8 degree of freedoms.



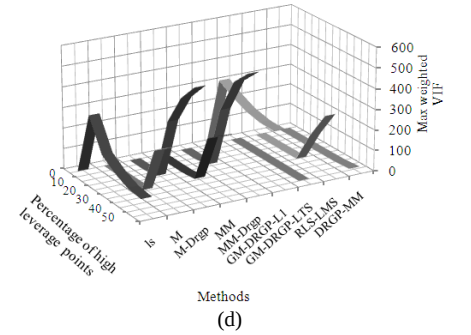
(a)



(b)



(c)



(d)

Fig. 1: Weighted multicollinearity diagnostics for sample size of 100 with three and seven explanatory variables; (a) n = 100, p = 3 (b) n = 100, p = 7 (c) n = 100, p = 3(d) n = 100, p = 7

Step 2: Rank the sum of the ranks for different error distribution in different number of explanatory variables. Table 6 presents the final Performance rank of AMSEE of different estimators for n = 100 in different distribution of error terms and different number of explanatory variables.

Step 3: Assign final rank (3) by ranking the sum of the final ranks for different explanatory variables. Table 7 consists of final Performance rank of AMSEE for different estimators in different number of explanatory variables where n = 100.

The maximum weighted correlation coefficient and maximum weighted VIF for sample of size 100 with three and seven explanatory variables are illustrated in Fig. 1. The results for five independent variables are consistent and are not included here due to space limitations. Here, a good estimator is the one in which the maximum correlation coefficient and the maximum weighted VIF are not easily affected by the presence of high leverage points.

DISCUSSION

Let us first focus our attention to the result of modified child mortality data set which is displayed in Table 2. The classical diagnostics measures of the original data clearly indicate that the data set doesn't have collinear explanatory variables. The T-tests and F-test confirm that there exists relationship between the explanatory and response variable. This data set has two multiple high leverage points based on the hat matrix by twice the mean-rule cutoff point, while DRGP (MVE) can detect 11 observations as multiple high leverage points. The residual standard error of the model is quite high due to the value of coefficient of determination (0.71)[8]. The high leverage points aren't collinearity-enhancing observations evident by the small value of correlation matrix and VIF (Table 2). The results of Table 3 signify that all the T'_2 of these multiple high leverage points for the original data exceeds the cutoff point of 3 which can be considered as high leverage points in X_2, except for observations 1, 5, 38 and 54. It is interesting to point out that after the modification (values for variable X_1 are modified to become high leverage collinearity-enhancing observations), the hat matrix can't detect all of these modified observations as multiple high leverage points while the DRGP (MVE) statistics identified them as high leverage points. The result of Table 2 suggests that there is a strong multicollinearity in the modified data set. Moreover, the non-significant of the t-statistics and the significant of the F-statistics of the two coefficient estimations confirmed the presence of multicollinearity in the modified data. The presence of multicollinearity has produced larger standard deviation of the errors for

the modified data as well. It is important to point out that the F-statistics for the DRGP-MM estimator as shown in Table 4 can't be obtained because it is not a one step reweighted estimator. It can be observed from Table 4 that among the proposed robust methods, only three estimators, that is the DRGP-MM, GM-DRGP-LTS and RLS-LMS can solve the multicollinearity problems. This result also suggests that the other methods can hardly rectify the multicollinearity problem evident by the larger p values and higher VIF values. It is interesting to note that the DRGP- MM has the least standard deviation error, followed by the GM-DRGP- LTS and RLS-LMS. We have not pursued the analysis of this example to the final conclusion, but a reasonable interpretation up to this stage is that the proposed Multi-stage GM-estimators and weighted MM-estimator which incorporated the DRGP are able to solve the problem of multicollinearity which is caused by high leverage points.

Next we will discuss the simulation results whether they confirm the conclusion of the numerical examples that our proposed methods performs better than the existing methods. It can be observed from Table 6 that DRGP- MM and GM-DRGP-LTS are equally good in the situation where the distribution of the error terms is normal. Based on the performance rank and final performance rank of AMSEE of Table 6 and 7, respectively, the DRGP -MM has the lowest final rank value followed by the GM-DRGP- LTS estimator. It is interesting to point out that several Multi-stage estimators, namely the MM-DRGP and M-DRGP are not performing better than one-stage estimators that is the RLS-LMS, MM-estimator and M-estimator. Thus, selecting different estimators to be used in each stage in the Multi-stage estimators are important issue to be considered.

Let us now focus to the result of Fig. 1. The plots in Fig. 1a and b show that the Maximum weighted correlation coefficient for LS method is equal to 1 which signify that the LS is very sensitive to high leverage points. Increasing the percentage of high leverage points in Fig. 1a and b, has increased the correlation coefficient of all methods except DRGP-MM and GM-DRGP-LTS. Moreover, the specific weights can't reduce the maximum weighted correlation coefficient much except these two new proposed methods. Any change in the number of explanatory variables changes the result slightly but still acceptable.

It can be seen from the maximum weighted VIF plots in Fig.1c and d that the maximum weighted VIF is less than the cutoff point of 10 for several estimators at low percentage of high leverage points. However, as the percentage of high leverage points increases, the maximum weighted VIF of most estimators exceed the cutoff point, except the DRGP-MM and GM-DRGP-LTS. It is important to mention here that when the percentage of high leverage points increases up to 20%, the maximum weighted VIF of LS method increases sharply and then decreases at a slower rate. However, the maximum weighted VIF values of LS method are still more than the cutoff point. In addition to that, by increasing the number of explanatory variables, the maximum weighted VIF of almost all of robust methods increases. For instance at 50% level of high leverage points, the maximum weighted VIF of LS method for $p = 3$ and $p = 7$ are equal to 42.47 and 77.52, respectively. The results of the maximum weighted VIF agree reasonably well with the results of the maximum weighted correlation coefficient and the preceding results that the two newly proposed methods outperform other methods considered in this study.

CONCLUSION

Outliers in the X-direction which are refer as multiple high leverage points can render least squares estimation meaningless and cause multicollinearity problems. Many robust methods have been developed to reduce the effect of outliers in the X-direction. Nonetheless, the development of robust methods that deal with the multicollinearity problems which are mainly due to multiple high leverage points has not been published extensively in the literature. The main focus of this study is to develop a reliable method for correcting the problem of high leverage points enhancing multicollinearity. In this study we incorporate the DRGP (MVE), one of the latest multiple high leverage diagnostics method with different types of robust estimators. The empirical study indicates that the DRGP-MM emerge to be more efficient and more reliable than other methods, followed by the GM-DRGP-LTS as they are able to reduce the most effect of multicollinearity. The results seem to suggest that the DRGP-MM and the GM-DRGP-LTS offers a substantial improvement over other methods for correcting the problems of high leverage points enhancing multicollinearity.

REFERENCES

1. Anderson, C. and R.E. Schumacker, 2003. A comparison of five robust regression methods with ordinary least squares: Relative efficiency, Bias and test of the null hypothesis. Understand. Stat., 2: 79-103. DOI: 10.1207/S15328031US0202_01
2. Andersen, R., 2008. Modern Methods for Robust Regression. Sage Publications, The United States of America, ISBN: 9781412940726, pp: 128.

3. Armstrong, R.D. and M.T. Kung, 1978. Least Absolute Values Estimates for a simple linear regression problem. J. R. Sci. Soc., 27: 363-366. http://www.jstor.org/pss/2347181
4. Beaton, A.E. and J.W. Tukey, 1974. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics, 16: 147-185. http://www.jstor.org/stable/1267936
5. Belsley, D.A., E. Kuh and R.E. Welsch, 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York. ISBN: 0471058564, pp: 292.
6. Belsley, D.A., 1991. Conditioning Diagnostics: Collinearity and Weak Data in Regression. Wiley, New York, ISBN: 10: 0471528897, pp: 396.
7. Coakley, C.W. and T.P. Hettmansperger, 1993. A bounded influence, high breakdown, efficient regression estimator. J. Am. Stat. Assoc., 88: 872-880. http://cat.inist.fr/?aModele=afficheN&cpsidt=3743502
8. Gujarati, D.N., 2002. Basic Econometrics. 4th Edn., Macgraw-Hill, New York, ISBN: 10: 0072478527, pp: 1002.
9. Habshah, M., M.R. Norazan and A.H.M.R. Imon, 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. J. Applied Stat., 36: 507-520. DOI: 10.1080/02664760802553463
10. Hadi, A.S., 1988. Diagnosing collinearity-influential observations. Comput. Stat. Data Anal., 7: 143-159. DOI: 10.1016/0167-9473(88)90089-8
11. Hadi, A.S., 1992. A new measure of overall potential influence in linear regression, Comput. Stat. Data Anal., 14: 1-27. DOI: 10. 1016/0167-9473(92)90078-T
12. Hampel, F.R., 1974. The influence curve and its role in robust estimation. J. Am. Stat. Assoc., 69: 383-393. http://www.jstor.org/pss/2285666
13. Hill, R.W., 1977. Robust regression when there are outliers in the carriers. Unpublished Ph.D. Dissertation, Harvard University, Boston, MA. http://proquest.umi.com/pqdweb?did=755042771&sid=3&Fmt=1&clientId=36652&RQT=309&VName=PQD
14. Hoaglin, D.C. and R.E. Welsch, 1978. The hat matrix in regression and ANOVA. Am. Stat. Assoc.,32:17-22. http://www.jstor.org/pss/2683469
15. Huber, P.J., 1973. Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Stat., 1: 799-821. http://www.jstor.org/stable/2958283
16. Imon, A.H.M.R., 2005. Identifying multiple influential observations in linear regression. J. Appl. Stat., 32(9): 929-946. DOI: 10.1080/02664760500163599
17. Kamruzzaman, M. and A.H.M.R. Imon, 2002. High leverage point: Another source of multicollinearity. Pak. J. Stat., 435-448. http://www.pakjs.com/journals/18(3)/18(3)7.pdf
18. Neter, J., M.H. Kutner, W. Wasserman and C.J. Nachtsheim, 2004. Applied Linear Regression Models. 3rd Edn., MacGraw-Hill, New York, ISBN: 10: 025608601X, pp: 720.
19. Marquardt, D.W., 1970. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. Technometrics, 12: 591-612. http://www.jstor.org/stable/1267205
20. Maronna, R.A. and V.J. Yohai, 2000. Robust regression with both continuous and categorical predictors. J. Stat. Plan. Inference, 89: 197-214. DOI: 10.1016/S0378-3758(99)00208-6
21. Maronna, R.A., R.D. Martin and V.J. Yohai, 2006. Robust Statistics: Theory and Methods. John Willy, New York, ISBN: 10: 0470010924, pp: 436.
22. Rosen, D.H., 1999. The diagnosis of collinearity: A Monte Carlo simulation study. Ph.D. Dissertation, Department of Epidemiology, School of Emory University, pp: 117. http://proquest.umi.com/pqdweb?did=730239851&sid=2&Fmt=2&clientId=36652&RQT=309&VName=PQD.
23. Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. Math. Stat. Appli., B: 283-297. http://www.ams.org/mathscinet-getitem?mr= 851060.
24. Rousseeuw, P.J., 1984. Least median of squares regression. J. Am. Stat. Assoc., 79: 871-880. http://www.jstor.org/stable/2288718
25. Rousseuw, P.J. and A.M. Leroy, 2003. Robust Regression and Outlier Detection. John Willy, New York, ISBN: 10: 0471852333, pp: 352.
26. Simpson, J.R., 1995. New methods and comparative evaluations for robust and biased-robust regression estimation. Ph.D. Dissertation, Arizona State University. http://www.stormingmedia.us/87/8758/A875892.html
27. Simpson, D.G., D. Ruppert and R.J. Carroll, 1992. On One-step GM estimates and stability of influences in linear regression. J. Am. Stat. Assoc., 87: 439-450. http://www.jstor.org/stable/2290275
28. Walker, E., 1985. Influence, collinearity and robust estimation in regression. Ph.D. Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia. http://proquest.umi.com/pqdweb?did=752570911&sid=1&Fmt=2&clientId=36652&RQT=309&VName=PQD.
29. Yohai, V.J., 1987. High breakdown-point and high efficiency robust estimates for regression. Ann. Stat. 15: 642-656. DOI: 10.1214/aos/1176350366