Journal of Mathematics and Statistics 7 (2): 129-136, 2011
ISSN 1549-3644

Estimation Methods for Multicollinearity Proplem Combined with High Leverage Data Points

Moawad El-Fallah and Abd El-Sallam
Department of Statistics and Mathematics and Insurance,
Faculty of Commerce, Zagazig University, Egypt

Abstract: Problem statement: Least Squares (LS) method has been the most popular method for estimating the parameters of a model due to its optimal properties and ease of computation. LS estimated regression may be seriously affected by multicollinearity which is a near linear dependency between two or more explanatory variables in the regression models. Although LS estimates are unbiased in the presence of multicollinearity, they will be imprecise with inflated standard errors of the estimated regression coefficients. **Approach:** In this study, we will study some alternative regression methods for estimating the regression parameters in the presence of multiple high leverage points which cause multicollinearity problem. These methods are mainly depend on a one step reweighted least square, where the initial weight functions were determined by the Diagnostic-Robust Generalized Potentials (DRGP). The proposed alternative methods in this study are called GM-DRGP-L_1, GM-DRGP-LTS, M-DRGP, MM-DRGP and DRGP-MM. **Results:** The empirical results of this study indicated 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:** The study had established that the DRGP-MM and the GM-DRGP-LTS methods were recommended to solve the multicollinearity problem with high leverage data points.

Key words: Multicollinearity problem, multiple high, leverage points, alternative robust estimations robust generalized , explanatory variables, diagnostic tool, regression analysis., high leverage, substantial improvement, leverage points, bounded influence

INTRODUCTION

Least squares estimation is one of the most important regression techniques used for estimating the parameters of a model. 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. The second 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. Hadi (1992) noted that collinearity-influential points are usually the points with high-leverage which tends to pull the model fit to their direction and introduced these points as a new source of multicollinearity problems. Thus, diagnosing the multiple high leverage points and recognizing estimations, methods which are resistant to these points may improve regression estimations. In this respect, alternative 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, alternative robust regression methods limit the influence of outliers. Three most important properties of any alternative robust regression method are efficiency, breakdown point and bounded influence (Andersen, 2008). The main objective of this study is to propose some alternative estimators that are able to perform well where multiple high leverage points cause multicollinearity problem in regression analysis. Nonetheless, the development of such estimators has not been published extensively in the literature. To

Corresponding Author: Moawad El-Fallah, Department of Statistics and Mathematics and Insurance, Faculty of Commerce, Zagazig University, Egypt

achieve this objective, different types of related alternative robust methods have been investigated and their properties are compared. Among the different types of robust techniques, we will consider the bounded influence or Generalized M-estimators (Marronna and Yohai, 2000; Ghazi *et al*., 2010; Ramzi. and Viviane, 2010) which attempt to assign less weight to the high influence observations and large residual points. To enhance the GM-estimators, these estimators may be defined as multi-stage estimators where in different stages, different alternative robust properties of each technique are applied to combine the desirable properties of each technique (Simpson *et al*., 1992).

MATERIALS AND METHODS

Robust regression methods: Let us consider the following linear regression model as Eq. 1:

$$Y = X\beta + \in \quad (1)$$

Where:
Y = The n × 1 vector of response
X = The n × P (P = k+1) matrix
$\in$ = The n × 1 vector which has standardized normal distribution

When the Least Squares (LS) method is employed, estimation of the regression parameters can be obtained from Eq. 2:

$$\hat{\beta} = (x / x)^{-1} x / Y \quad (2)$$

Robust regression procedures are mainly aim to provide resistant (stable) results in the presence of outliers. The Least Absolute Values (LAV) is one of the first robust methods that was introduced by Armstrong and Kung (1987) with a higher efficiency than LS by minimizing the sum of the absolute residuals. The use of this criterion, rather than ordinary least squares, provides robustness against outliers and is particularly useful when the $\in_i$ disturbances are generated by fat-tailed distributions. Rousseuw (1984). Rousseuw *et al*. (2003) introduced two robust methods namely the Least Median of Squares (LMS) and the Least Trimmed Squares (LTS). The LMS attempts to minimize the median of e^2_i while the LTS minimize $\sum_{i=1}^{h} e^2_{(i)}$ where $e^2_{(1)} \leq e^2_{(2)} \leq, \ldots, \leq e^2_{(n)}$ are the ordered squared residuals, i =1,…,n and h is the number of residuals included in the calculation. Both estimators have high breakdown, that 50%. However, they are unbounded influence estimators, where the LMS and the LTS has low and medium efficiency value, respectively (Simpson *et al*. (1992). Huber (1973) proposed a robust M-estimators where $\hat{\beta}$ are obtained by solving Eq. 3:

$$\sum_{i=1}^{n} \psi\left(\frac{yi - x_{i'}\hat{\beta}}{s}\right) \quad (3)$$

It is important to point out that, there are two types of ψ-functions, that is the monotonic ψ-functions (e.g. Huber's ψ-functions) and the redescending ψ-functions (e.g., biweight ψ function's, Beaton and Tukey (1974). M-estimators are the simplest high efficiency robust procedures, both computationally and theoretically having desirable asymptotic properties. However, the M-estimator is not robust in the X-direction and has a low breakdown point, that is equal to (1/n) (Simpson *et al*., 1992; Marronna *et al*., 2006) introduced a class of methods which is called the Generalized M-estimators (GM-estimators) with a major aim of downweighting those high leverage points which have large residuals. Marronna *et al*. (2006) also, reported that these estimators have high efficiency and bounded influence properties which achieve a moderate breakdown point equal to (1/P).

The GM-estimator is the solution of the normal equation (4):

$$\sum_{i=1}^{n} \varphi_i \, \psi\left(\frac{y_i - x_i \hat{\beta}}{s\varphi =}\right) x_1 = 0 \quad (4)$$

where, φ are defined to down weight high leverage points, with high residuals and s is a robust scale estimate. Iteratively Reweighted Least Squares (IRLS) may be used to solve (4). At convergence, the GM-estimator may be written as Eq. 5:

$$\hat{\beta}_{GM} = \left(X'WX\right)^{-1} X'WY \quad (5)$$

where, in this case, the diagonal elements of W are the weights w_i defined as Eq. 6:

$$W_i = \frac{\psi\left[\frac{y_i - x_i \hat{\beta}_{GM}}{\varphi_i s}\right]}{(y_i - x_i \hat{\square}_{GM})\varphi_i s} \quad (6)$$

The main objective of this study is to study some alternative estimators that are able to perform well where multiple high leverage points are the cause of the multcollinearity problem in regression analysis. In particular, 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 with a major aim of down weighting the high leverage points which have large residuals. Hence, in this study, we propose mainly alternative 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. Unfortunately, the MM-estimators are also sensitive to outliers in X-variables. As a solution to this drawback of MM-estimators, an alternative robust method is developed in section four.

To confirm the advantage of our alternative proposed methods, these methods compared with reweighted least square based on LMS (RLS-LMS) defined by Rousseuw and Leroy (2003). They computed scale estimator as:

$$s^0 = 1.4826 * \left(1 - \frac{5}{n-p-1}\right) * \sqrt{med(r_i^2)}$$

where, r_i is the residual of LS. The following hard rejection function for standardized residuals $\frac{r_i}{S}$ is utilized to compute the following initial weights Eq. 7:

$$w_i = \begin{bmatrix} 1 & \text{if } \left|\frac{ri}{S^0}\right| \le 2.5 \\ 0 & \text{otherwise} \end{bmatrix} \quad (7)$$

The final weights for RLS-LMS are identified by the usage of a hard rejection for standardizing the LMS residuals by new scale as scale $= \sqrt{\frac{w * r^2}{\sum_{i=1}^{n} w_i - p - 1}}$ the weakness of this method is using LMS which is a low efficiency estimator.

Diagnostic robust generalized potential statistics: 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 Eq. 8:

$$T = \frac{x - \bar{x}}{S} \quad (8)$$

Where:

$\bar{x}$ = The mean

s = The standard deviation of collinear explanatory variables

The robust version of (8) is Eq. 9:

$$T' = \frac{x - median(x)}{Max(x)} \quad (9)$$

where, Mad (x) is the normalized median absolute deviation about the Median (x) (Mad= 1.4826(median $|r_i$-median $(r_i)|$). 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 (Marronna *et al*., 2006). This method can be used in univariate regression models as a diagnostics rule to detect high leverage points. Since in most of the regression analysis, more than one collinear 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 $W=X (X^TX)^{-1}X^T$. The most widely used cutoff points of hat matrix is twice-the- mean-rule (2k/n) by Hoaglin and Welsch (1978). Hadi (1988) 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. He introduced another diagnostic tool as follows Eq. 10:

$$p_{ii} = \frac{w_{ii}}{1 - w_{ii}} \quad (10)$$

where, $w_{i\,i} = x_i^T \left(x^T x\right) - \bar{x}_i$ is the diagonal element of W and the i-th, diagonal potential p_{ii} can be defined as $p_{ii} = x_i^T \left(x_{(i)}^T x_{(i)}\right)^{-1} x_i$, where $X_{(i)}$ is the data matrix X without the i-th row. He proposed a cut off point for potential values p_{ii} as Median (p_{ii}) + c Mad (p_{ii}) (MAD-cutoff point) and c can be taken as constant values of 2 or 3. This method also is unable to detect all of the high leverage points. So, Hadi (1988) introduced another diagnostic tool as generalized potentials for the whole data set which is defined as Eq. 11:

$$p_{ii}^* = \begin{cases} \frac{w_{ii}^{(-D)}}{1 - w_i^{(-D)}} & \text{for } i \in R \\ w_{ii}^{(-)} & \text{for } i \in D \end{cases} \quad (11)$$

Where:

D = The deleted set which corresponds to the suspected outliers

R = The remaining set from observations after deleting d < (n-k) and it 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*. (2009) 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) (RMD-MVE) (defined by Rousseuw (1985) as Eq. 12:

$$\begin{aligned} &RMD_i = \sqrt{(x - T_R(x))' C_R(x)^{-1} (□ - T_R(x))} \\ &\text{for } i = 1,...,n \end{aligned} \tag{12}$$

where, T_R (×) and C_R (×) are robust location and shape estimate such as MCD or MVE. The RMD- MVE has been used to detect the suspected group (D group) in generalized potential method in (11). The merit of this method is swamping less good leverage as high leverage points comparing with the RMD -MVE. In the next section we propose robust methods based on DRGP.

Alternative proposed robust methods: In this section, some proposed form of φ-weights in (4) are generated and discussed. It is important to point out that in the proposed methods, P_i is the DRGP statistics with MAD-cutoff of this statistic. In these methods, we will employ the Tukey's biweight redescending φ-function which is defined as Eq. 13:

$$\psi(t) = \begin{cases} t\left[1 - \left(\frac{t}{c}\right)^2\right]^2 & \text{if } |t| \le c \\ & \text{if } |t| > c \end{cases} \tag{13}$$

The Tukey's biweight with the tuning constant c = 4.685 will result a 95% efficiency under normal error distribution. Assigning lower weights (even zero if the residual is too large) to large outliers, a redescending φ-function is better compared to monotonic functions such as Huber's function. In this respect, a redescending φ-functions limits the influence of outliers more effectively than a monotone φ-function. Given the fact that iterative techniques are computationally expensive and there is no guaranty to result in better estimations (Simpson, 1995) a one step reweighted LS is used for most of the proposed methods except DRGP-MM estimator. Simpson *et al*. (1992) enumerated that the GM and MM-estimators surpass other robust method. In this connection, most of the alternative proposed methods are similar to that of GM and MM estimators with alight modification in which the DRGP proposed by Habshah *et al*. (2009) is incorporated in the calculation of the φ. The first two proposed estimators are multi-stage GM-estimators, while the others are defined based on the M-estimator and MM-estimators. The proposed methods will be computed in three steps and summarized as follow.

GM-DRGP-L_1:

Step 1: Employ L_1 as initial estimate and then obtain the standardized residuals of L_1 estimator. Compute MAD = 1.4826 (med|r_i-med (r_i)|. according to Marronna and Yohai (2000). 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: Defining $\varphi_{\downarrow}i = \min\left[1, \left(\text{MAD} - \text{cutoff}\left(p_{\downarrow}i\right)\right) / \left(\left(p_{\downarrow}i\right)\right)\right]$ in (4) and using function (13) 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 $\varphi_{\downarrow}i = \min\left[1, \left(\text{MAD} - \text{cutoff}\left(p_{\downarrow}i\right)\right) / \left(\left(p_{\downarrow}i\right)\right)\right]$ in (4) and using function (13) to assign final weights to the observations.

Step3: Compute a one step reweighted least squares as convergence approach.

M-DRGP:

Step 1: Compute the residuals of M-estimates of scale by assigning the initial weight of W_i, (DRGP (MVE) $\min\left[1, \left(\text{MAD} - \text{cutoff}\left(p_{\downarrow}i\right)\right) / \left(\left(p_{\downarrow}i\right)\right)\right]$ where P_i is DRGP (MVE) statistics.

Step 2: Define new weights as $w_i = r_i$ (M-estimator)/scale (M-estimator) and using 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 is replaced with the MM-estimators.

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 (13) 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 (1985) noted that sometimes the weighting process in robust methods can improve 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). 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 (Rosen, 1999). Since, this multicollinearity diagnostics is simple and easy to compute, it is more preferred (Belsley, 1991, Belsley *et al*., 1990). Another practical approach to detect multicollinearity is by using variance inflation factors (VIF). VIF is defined as V1F (i) $\left(1-R_i^2(w)\right)^{-1}$ 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 (Marquardt, 1970). A cutoff point of (11) is recommended as a rule of thumb for VIF to detect severe multicollinearity. The weighted linear regression can be expressed as a transformed model Eq. 14:

$$Y_w = X_{w\beta} + \epsilon_w \quad (14)$$

where, $Y_w = W^{1/2}Y$, $X_w = W^{1/2}X$ and $\epsilon_w = W^{1/2}\epsilon$ (Neter *et al*. (2004). The final weight of the proposed estimator, which is 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 that are responsible for the multicollinearity. It is important to point out that all high leverage points are not collinearity-influential and vice versa (Hadi, 1992) The weighted correlation matrix can be computed through the correlation matrix of X_w. The weighted VIF is defind as follows Eq. 15:

$$VIF_w(i) = \left(1-R_i^2(w)\right)^{-1} \quad (15)$$

where, R^2 (W) is the coefficient of determination of regressing each X_{wi} on the other weighted explanatory variables. It is worth to mention 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: In this section we consider a real data set to evaluate the performance of our proposed robust methods.

Child mortality data set: Gujarati (2002) 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 1 presents the classical multicollinearity diagnostics methods such as the correlation matrix and VIF. The classical diagnostics measures of the original data clearly indicates that the data set doesn't have collinear explanatory variables. The T and F-tests confirm that there exists relationship between the explanatory and response variable. This data set has two high leverage points based on the hat matrix by twice-mean-rule cutoff point, while DRGP (MVE) can detect 11 observations as high leverage points.

Table 1: Multico llinearity diagnostics and least square coefficients of 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	1.08	-2.23 (0.007)	-0.01 (0.000)	0.000	41.75
Modified data	0.99	37.34	-0.24 (0.512)	0.002 (0.938)	0.003	70.25

Table 2: High leverage diagnostics for child mortality data set

Original data					Modified data			
index	Hat(x) (0.09)	DRGP(x)(0.11)	$T_1(3)$	$T_2(3)$	index	modified X_1	Hat(x) (0.09)	DRGP(x)(0.11
1	0.02	0.20	0.34	2.16	24	248	0.03	1.23
5	0.03	0.14	0.89	2.47	27	180	0.02	0.55
24	0.05	0.90	1.03	6.26	30	1107	0.75	34.72
27	0.05	0.35	1.22	4.15	33	490	0.13	6.04
30	0.77	31.67	0.47	33.22	53	258	0.04	1.36
33	0.14	5.52	0.06	13,87	58	255	0.11	0.14
38	0.05	0.15	1.25	2.54	62	214	0.03	0.85
53	0.05	102	0.97	6.59				
54	0.05	014	1.10	0.60				
58	0.07	0.91	1.47	6.48				
62	0.05	0.59	1.16	5.21				

Table 3: Multicollinearity diagnostics and least square coefficients of different methods for modified child mortality data set

Method	Cor (x_1,x_2)	VIF	b_1(t p-value)	b_2(t p-value)	F p-value	s (e)
Ls	0.99	37.34	-0.24	0.002	0.00	70.25
			(0.512)	(0.938)		
GM-DRGP-L_1	0.98	33.36	-0.68	-0.02	0.00	55.00
			(0.02)	(0.34)		
GM-DRGP- LTS	0.65	1.75	-1.78	-0.04	0.00	39.20
			(0.00)	(0.00)		
DRGP- MM	0.65	1.74	-1.61	-0.04		36.17
			(0.00)	(0.00)	-	-
MM-DRGP	(0.02)	0.90	-0.69	-0.01		
			(0.02)	(0.33)	0.00	54.80
M-DRGP	0.90	5.25	-0.69	-0.01	0.00	54.80
			(0.02)	(0.33)		
RLS-LMS	0.90	5.25	-1.80	-0.04	0.00	39.45
	0.65	1.74	(0.00)	(0.00)		

The high leverage points aren't collinearity-enhancing observations evident by the small value of correlation matrix and VIF (Table 1). It is important to note here that the high leverage points will be the prime source of multicollinearity when they are in the same observations with at least two explanatory variables. The robust three-sigma edit rule (Eq. 9) is shown in Table 2. The results of Table 2 signify that all the T '$_2$ exceeds the cutoff point of 3 which can be considered as high leverage points, except observations 1,5,38,and 54. In order to obtain a large magnitude of high leverage points in x_1 as in x_2 , a modification in x_2 in the points 24, 27, 30, 33, 53, 58- 62 has been considered based on the following formula:

$$\text{Modified}\left(x_1\right) = T_2^*\left(\text{mad}\left[\!\left[(x\right]\!\right]_1\right) + \text{median}\left(x_1\right)$$

The modified x_1 is also displayed in Table 2. It is interesting to point out that after the modification, the hat matrix can't detect all of these modified observations as high leverage points, while the DRGP (MVE) statistics identified them as high leverage points. The results of Table 1 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 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.

Table 3 presents the multicollinearity diagnostics and least squares coefficients of the modified child mortality data set for proposed robust methods and the existing robust methods which were introduced previously.

The results of Table 3 point out that the F-statistics can't be obtained for DRGP- MM estimator because it is not a one step reweighted estimator. It can be shown also from Table 3 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. It is interesting to note that the DRGP- MM has the least standard deviation error, followed by the GM-DRGP- LTS and RLS-LMS.Thus the new proposed estimators namely the DRGP-MM and the GM-DRGP-LTS, outperforms all other defined estimators.

DISCUSSION

Let us first focus our attention to the result of modified child mortality data set which is displayed in Table 1. The classical diagnostics measures of the original data clearly indicate that the data set does not have collinear explanatory variables. The T and F- tests 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 high leverage points are not collinearity-enhansing observations evident by the small value of correlation matrix and VIF (Table 1). The results of Table 2 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 observation 1,5,38and 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 not 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 1 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 3 can not be obtained because it is not a one step reweighted estimator. It can be observed from Table 3 that, among the proposed robust methods, only three estimators, that is the DRGP-MM, GM-DRGP-LTS and RLT-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-incorporated the DRGP are able to solve the problem of multicollinearity which is caused by high leverage points.

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 offers a substantial improvement over other methods for correcting the problems of high leverage points enhancing multicollinearity.

REFERENCES

Andersen, R., 2008. Modern Methods for Robust Regression. 1st Edn., Stage Publication, The United States of America, ISBN: 9781412940726, pp: 107.

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.

Beaton, A.E. and J.W. Tukey, 1974. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics, 16: 147-185.

Belsley, D.A., 1991.Conditioning Diagnostics: Collinearity and Weak DAT in Regression. 1st Edn., Wiley, New York, ISBN: 10: 0471528897, pp: 396.

Belsley, D.A., E. Kuh and R.E. Welsch, 1980. Regression Diagnostics: Identing Influential Data and Sources of Collinearity. 1st Edn., Wiley. New York, ISBN: 0471058564, pp: 292.

Ghazi, F.M., A. Majed, Alsharayri and M.D. Mwafaq, 2010. Impact of firm’s characteristics on determining the financial structure on the insurance sector firms in jordan. J. Soc. Sci., 6: 282-286. DOI: 10.3844/jssp.2010.282.286

Gujarati, D.N., 2002. Basic Econometrics. 4th Edn., Macgraw-Hill , New York, ISBN: 10: 0072478527, pp: 1002.

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: 10.11080/02664760802553463

Hadi, A.S., 1988. Diagnosing collinearity-influential observations. Comput. Statis. Data Anal., 7: 143-159. DOI: 10.1016/0167-9473(88)90089-8

Hadi, A.S.1, 1992. A new measure of overall potential influence in linear regression. Comput. Statis. Data Analysis, 14: 1-27. DOI: 10.1016/0167-9473(92)90078-T

Hoaglin, D.C. and R.E. Welsch, 1978. The hat matrix in regression and ANOVA. Am. Stat. Assoc., 32: 17-22.

Huber, P.J., 1973. Robust Regression: Asymptotic, conjectures and monte carlo. Annals o Statis., 1: 799-821.

Marquardt, D.W., 1970. Generalized inverses ridge regression, biased linear estimation and nonlinear estimation. Technimetrics, 12: 591-612.

Marronna, R.A. and V.J. Yohai, 2000. Robust regression with both continuous and categorical predictors. J. Stat. Plan. Inferen., 89: 197-214. DOI: 10.1016/S0378-3758(99)00208-6

Marronna, R.A., R.D. Martin and V.J. Yohai, 2006. Robust Statistics: Theory and Methods. 1st Edn., John Willy, New York, ISBN: 10: 0470010924, pp: 403.

Neter, J., M.H. Kutner, W. Wasserman and C.J. Nachtsheim, 2004. Applied Linear Regression Models. 3rd Edn., Macgraw-Hill, New York, ISBN: 10: 025608601 X, pp: 720.

Ramzi, N.N. and N. Viviane, 2010. Using regression to establish weights for a set of composite equations through a numerical analysis approach: A case of admission criteria to a college. J. Math. Stat., 6: 300-305. DOI: 10.3844/jmssp.2010.300.305

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/ d = pqdweb? Did = 730239851 and sid = 2 and Fmt = 2 and clientld = 36652 and RQT=309 and VName = PQD

Rousseuw, P.J. and A.M. Leroy, 2003. Robust Regression and Outlier Detection. 1st Edn., John Willy, New York, ISBN: 0471488550, pp: 329.

Rousseuw, P.J., 1984. Least median of squares regression. J. Am. Stat. Assoc., 79: 871-880.

Rousseuw, P.J., 1985. Multivariate estimation with high breakdown point. Math. Stat. Applied, B: 283-297..

Simpson, J.R., 1995. New methods and comparative evaluations for robust and biased-robust regression estimation. Ph. Dissertation, Arizona State University. http://www.stormingmedia.us/87/8758/A875892.

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.

Walker, E., 1985. Influence, collinearity and robust estimation in regression. Ph.D. Dissertation, Virginina Polytechnic Institute and State University, Blacksburg, Verginia. http://proquest.unit.com/pqdweb?Did=752570911 &sid=1&Fmt=2&clientId=36652&ROT=309&VN ame=PQD