Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 10 (8): 611-619, 2010
ISSN 1812-5654

The Application of Robust Multicollinearity Diagnostic Method Based on Robust Coefficient Determination to a Non-Collinear Data

[1]H. Midi, [1]A. Bagheri and [2]A.H.M.R. Imon
[1]Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2]Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, USA

Abstract: In this study, we proposed Robust Variance Inflation Factors (RVIFs) in the detection of multicollinearity due to the high leverage points or extreme outliers in the X-direction. The computation of RVIFs is based on robust coefficient determinations which we called RR^2 (MM) and RR^2 (GM (DRGP)). The RR^2 (MM) is coefficient determination of high breakdown point and efficient MM-estimators whereas RR^2 (GM (DRGP)) has been defined through an improved GM-estimators. The GM (DRGP) is a GM-estimator with the main aim as downweighting high leverage points with large residuals. It has been introduced by employing S-estimators as initial values, Diagnostic Robust Generalized Potential based on MVE (DRGP (MVE)) as initial weight function and an Iteratively Reweighted Least Squares (IRLS) has been utilized as a convergence method. The numerical results and Monte Carlo simulation study indicate that the proposed RVIFs are very resistant to the high leverage points and unable to detect the multicollinearity in the data especially RR^2 (GM (DRGP)). Hence, this indicates that the high leverage points are the source of multicollinearity.

Key words: Coefficient determination, generalized M-estimators, high leverage points, variance inflation factor

INTRODUCTION

When two or more independent variables of a linear regression are highly correlated with each other, multicollinearity is said to exist. Multicollinearity produces unexpectedly large standard errors of the Ordinary Least Squares (OLS) estimates. The presence of multicollinearity in the data set is suggested by non-significant results in individual tests on the regression coefficients for important explanatory variables where in fact they are significant. Since, multicollinearity causes major interpretive problems in regression analysis, it is very crucial to investigate and detect its presence to reduce its destructive effects on the regression estimates. There are different primary sources of multicollinearity, such as the data collection method employed, constraints on the model or in the population being sampled, model specification and an over determined model (Montgomery *et al*., 2001). It is now evident that high leverage points which fall far from the majority of the explanatory variables are another source of multicollinearity (Kamruzzaman and Imon, 2002). These points are considered as good or bad leverage points according to whether they follow the same regression line as the other data in the data set or not. Furthermore, collinearity influential observations are the observations which can change the collinearity pattern of the data. They may be enhancing or reducing collinearity in the data set. All the high leverage points are not collinearity influential observations and vice versa (Hadi, 1988). While large magnitude of high leverage points which exist in more than one explanatory variable may be collinearity-enhancing observations and they may bring collinearity for explanatory variables in non-collinear data sets. Among different multicollinearity diagnostics which exists in the literatures, the Variance Inflation Factors (VIF) is the commonly used method (Montgomery *et al*., 2001; Belsley *et al*., 1980) which is sensitive to the presence of high leverage points. It is worth mentioning that the high leverage points which is pointed by Kamruzzaman and Imon (2002) as a new source of multicollinearity in non-collinear data set, is based on the classical diagnostics method. It is important to note that the estimation of regression parameters is unbiased in the presence of multicollinearity (Montgomery *et al*., 2001; Belsley *et al*., 1980). However, when high leverage points cause collinearity in the data set, the coefficient estimations will become bias and this situation is not desired (Bagheri and Midi, 2009). Thus, it is imperative to investigate the source of collinearity in

Corresponding Author: Arezoo Bagheri, Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia

the data set since, the remedial measures to solve the problem of multicollinearity are highly dependent on the understanding of the differences among these sources of multicollinearity (Montgomery *et al*., 2001). The use of classical diagnostic methods for the detection of multicollinearity caused by high leverage points may produce a misleading conclusion. In collinear or non-collinear data sets, these points have different effect on the classical diagnostics methods. Sometimes the classical methods in the presence of high leverage points indicate non-collinearity for the collinear data set or collinearity for the non-collinear data set. Therefore, if the high leverage points are the only source of multicollinearity in the data set, robust methods can be employed to remedy this problem. In this situation, other remedial measures of multicollinearity problem such as the Principal Component Regression (PCR) (Jolliffe, 1982) are not necessary. Unfortunately, little work has been explored in the effect of high leverage points on the classical multicollinearity diagnostics method. In the presence of high leverage collinearity influential observations, the classical multicollinearity diagnostics methods such as the VIF diagnose the existence of collinearity in the data set but it doesn't reveal the source of multicollinearity. Making the VIF resistant to these high leverage points will guide us to the detection of the source of multicollinearity in the data set. High leverage points are the only source of multicollinearity if the Robust VIF doesn't diagnose multicollinearity in their presence. However, on the contrary, the VIF suggests that multicollinearity exists in the data set. Subsequently, the RVIF is unable to detect the existence of mlticollinearity. For a through overview of the robust methods, one can refer to Rousseeuw and Leroy (2003), Wilcox (2005), Maronna *et al*. (2006) and Andersen (2008). It is important to point out that the formulation of VIF is based on the coefficient determination of fitted regression line when each of the explanatory variables is regressed with the remaining predictor variables using the Ordinary Least Squares (OLS) Method. Nevertheless, it is now evident that outliers in the X or Y direction have an unduly effect on the OLS estimates (Midi *et al*., 2009) and subsequently the classical VIF. This situation has motivated us to develop new robust VIFs which are based on Robust coefficient determination (RR^2). Hence, the Robust coefficient determination (RR^2) has to be formulated first, prior to propose any robust VIF. Several robust coefficient determinations exist in the literature of robust methods such as Rousseeuw and Hubert (1997) and Splus 6 Robust Library User's Guide (2001). The Generalized M-estimators (GM-estimators) is one of the robust methods which have desirable properties that attempts to downweight high leverage points as well as large residuals. In this study, prior to developing a robust VIF, a new GM-estimator is proposed. The proposed method incorporates S-estimators defined by Rousseeuw and Yohai (1984) and DRGP (MVE) which is introduced by Midi *et al*. (2009). A robust coefficient determination is then proposed based on applying the new GM estimator to fit the regression model. Following this, the robust VIF is developed. Moreover, another Robust VIF is proposed by employing robust coefficient determination based on MM-estimator which is introduced by Yohai (1987). Our new proposed Robust VIFs will be applied to a well-known non-collinear data set. To compare the performance of these robust multicollinearity diagnostics methods, a Monte Carlo simulation study will also be carried out.

MATERIALS AND METHODS

New Proposed robust estimator: The multiple regression model can be expressed as:

$$y = x\beta+\varepsilon \tag{1}$$

where, Y is an n×1 vector of response or dependent variables, X is an n×p (n>k) matrix of predictors (explanatory variables), β is a p×1 vector of unknown finite parameters to be estimated and is an n×1 vector of random errors. When the Ordinary Least Squares (OLS) method is employed to estimate the regression parameters we obtain:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2}$$

In the presence of outliers in the X- or Y-directions, the OLS estimations are not reliable. Thus, utilizing robust methods will be necessary. The robust estimators through downweighting the outliers aim to fit the model according to the majority of the data. One of the resistant robust methods against high leverage points are GM-estimators (Hill, 1977) which is introduced by Schweppe. The major aim of these methods is to downweight those high leverage points which have large residuals or bad leverage points. These estimators have high efficiency and bounded influence properties which achieve a moderate break down point equal to 1/p (Simpson, 1995). The GM-estimator is the solution of the normal equation:

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{y_i - x\hat{\beta}}{s\pi_i}\right)x_i = 0 \tag{3}$$

where, π_i is defined to downweight high leverage points with high residuals, s is a robust scale estimate and ψ-function may be a monotonic ψ-function such as Huber's ψ-function which is defined as:

$$\psi(t) = \begin{Bmatrix} t & \text{if} \quad |t| < k \\ k\,\text{sgn}(t) & \text{if} \quad |t| \geq k \end{Bmatrix} \tag{4}$$

It is noticeable that k = 1.345 has been chosen to achieve 95% efficiency under normal error distribution. Iteratively Reweighted Least Squares (IRLS) may be used to solve Eq. 4. At convergence, the GM-estimator is written as follows:

$$\hat{\beta}_{GM} = (X'WX)^{-1}X'Wy \tag{5}$$

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

$$w_i = \frac{\psi\left[(y_i - x_i'\hat{\beta}_{GM})/\pi_i s\right]}{(y_i - x_i'\hat{\beta}_{GM})/\pi_i s} \tag{6}$$

Multi-stage GM-estimators have been developed in order to overcome the weakness of GM-estimators, that is low break down point. These estimators may have high break down point if we obtain appropriate initial estimators. One of the first GM-estimator with high efficiency, high breakdown (50%) and bounded influence was proposed by Coakley and Hettmansperger (1993) and (Wilcox, 2005). This method incorporated two most practical high breakdown robust estimators, that is the Least Median of Squares (LMS) and the Least Trimmed Squares (LTS). The LTS estimator is used as initial estimator and the LMS estimator is integrated in the development of scale estimate (Rousseeuw, 1984). The Robust Mahalanobis Distance (RMD) based on Minimum Volume Estimator (MVE) (Rousseeuw and van Zomeren, 1990) (RMD-MVE) also defined as leverage estimates. Rousseeuw (1985) defined RMD-MVE as:

$$RMD_i = \sqrt{(X - T_R(X))'C_R(X)^{-1}(X - T_R(X))} \tag{7}$$

where, T_R (X) and C_R (X) are robust location and shape estimate of MVE. In this estimator π-weight is a ratio of the χ^2 cutoff value to the squared Robust Mahalanobis Distance. A one step Newton Raphson has been used as convergence approach (for more details, one can refer to Wilcox (2005).

One of the drawbacks of this estimator is in the definition of π-weight which depends on RMD-MVE. The RMD-MVE 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 (Midi *et al.*, 2009; Imon *et al.*, 2009). Improving the precision of these estimators requires an effective diagnostics method that will identify appropriate π–weight. Simpson (1995) discussed several types of Multi-stage GM-estimators and a comparison evaluation of the existing robust estimators. In this study, a new GM-estimator whose major aim is downweighting those high leverage points which have large residuals will be utilized in developing our proposed method. The new GM-estimator which we called GM (DRGP)-estimator is proposed by employing the S-estimator as initial estimator instead of the LTS estimator in the GM-estimator algorithm of Coakley and Hettmansperger (1993). It has been verified that the asymptotic efficiency of this estimator is high (Andersen, 2008) for studying more about S-estimator and its properties. To overcome the shortcoming of π-weight in their algorithm, we propose to employ Diagnostic Robust Generalized Potential based on MVE (DRGP (MVE)) which is proposed by Midi *et al.* (2009). This latest diagnostics method of high leverage points has an attractive feature whereby it is able to identify the exact number of high leverage points that exist in the data set. The DRGP (MVE) can be defined as:

$$p_i = \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \qquad \text{for } i \in R \tag{8}$$

$$= w_{ii}^{(-D)} \quad \text{for } i \in D$$

where, R $\{i;\ RMD_i^2, <\text{MAD-cutoff}(RMD)\}^2$, a non-parametric cutoff point which is called MAD-cutoff and defined as MAD-cutoff(θ) = Median(θ)+K*Mad(θ) where, K is set to be the constant values of 2 or 3. The RMD_i^2 is introduced in Eq. 7. Furthermore, D = R^T and R^T indicates the complement of collection R.The merit of this method is swamping less good leverage as high leverage points compared to the RMD-MVE.

Hence, the proposed algorithm for finding GM (DRGP)-estimator is as follow: To compute a GM estimator, begin by setting k = 0 and computing the S-estimate of the intercept and slope parameters, $\hat{\beta}_{0k},\ldots,\hat{\beta}_{Pk}$. Proceed as follows:

Step 1: Compute the residuals of the S-estimator as initial estimator $r_{i,k} = y_i - \hat{\beta}_{0k} - \hat{\beta}_{1k}x_{i1} - \ldots - \hat{\beta}_{Pk}x_{il}$ and scale the residuals by applying:

$$\hat{\tau}_k = 1.4826(1 + 5/(n-p))\text{Median}|r_{ik}|$$

Step 2: Form π-function as:

$$\pi_i = \min\left[1, \frac{MAD - cutoff(p_i)}{(p_i)}\right] \quad i=1,....,n$$

where, p_i is defined in Eq. 8.

Step 3: Define the initial weights as:

$$w_i = \frac{\hat{\tau}\times\pi_i}{r_i}\psi\left(\frac{r_i}{\hat{\tau}\times\pi_i}\right)$$

for i = 1,…,n where, a Huber's ψ-function which has been introduced in Eq. 4 is applied.

Step 4: Use these weights to obtain a weighted least squares estimates, $\hat{\beta}_{0k+1},...,\hat{\beta}_{Pk+1}$. Increase k by step 1

Step 5: Repeat steps 1-4 until convergence. That is, iterate until the change in the estimated parameters is small

Another high break down point estimator which has been defined by Rousseeuw and Yohai (1984) is called S-estimators. These estimators are the solution that finds the smallest possible dispersion of the residuals:

$$\min \hat{\sigma}(e_1(\hat{\beta}),...,e_1(\hat{\beta})) \tag{9}$$

Instead of minimizing the variance of the residuals; this robust S-estimation minimizes a robust M-estimate of the residual scale:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{\hat{\sigma}_e}\right) = b \tag{10}$$

where, b is a constant defined as b = Eϕ[ρ(e)] and ϕ represents the standard normal distribution. Differentiating Eq. 8 and solving the following resultant equation:

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{e_i}{\hat{\sigma}_e}\right) = b \tag{11}$$

where, ψ is an appropriate weight function. The S-estimates have high efficiency which is more than the LTS estimators relative to the OLS (Croux *et al*., 1994) and has high breakdown point of 50%.

It is important to note that the MM-estimator is one of the most important robust estimators which is first proposed by Yohai (1987). These estimators combine high breakdown value estimators (50%) and M-estimators which have high efficiency (approximately 95% relative to OLS under the Gauss-Markov assumptions). The MM-estimators in the name refers to the fact that more than one M-estimation procedure is used to calculate the final estimates.

Robust Variance Inflation Factor: Marquardt (1970) proposed the most popular diagnostic tools of multicollinearity, namely, the Variance Inflation Factor (VIF) which measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. Thus it is define as follows:

$$VIF_j = \frac{1}{1-R_j^2} \quad j=1,....,k \tag{12}$$

where, R^2 is the coefficient determination of each of the explanatory variables when regressed on the other explanatory variable in the ordinary regression model by using the OLS method. Moderate collineriaty exists in the data set when VIF is between 5 and 10. Any VIF value that exceeds its cutoff point 10, indicates that the associated regression coefficients are poorly estimated because of severe multicollinearity.

This multicollinearity diagnostics method is highly sensitive to the presence of high leverage points because of the effect of these points on R^2. Consequently, misleading conclusions are obtained from the classical VIF. In this respect, it is imperative to formulate a robust diagnostics method to avoid from making a wrong conclusion. Since, the computation of VIF is highly dependent on the calculation of R^2, the robust version of VIF also can be defined from RR^2. In this study, we develop RVIF which are based on two robust coefficient determinations, namely the RR^2 = (MM) and the RR^2 (GM(DRGP).

The RR^2 (MM) is one of the handiest Robust coefficient determination (RR^2) and can be obtained from robust library of SPLUS software. It can be calculated as follows:

If the corresponding coefficient estimates are the initial S-estimates (RR^2), the RR^2 (MM) is computed using the initial S-estimates. If an intercept term is included in the model, then the RR^2 (MM) is defined as:

$$RR^2(MM) = \frac{(n-1)s_y^2 - (n-p)s_e^2}{(n-1)s_y^2} \tag{13}$$

where, $s_e = \hat{s}^0$ and s^y is the minimized $\hat{s}(\mu)$, for a regression model with only an intercept term with parameter μ. If the corresponding coefficient estimates computed using the final M-estimates, it can be obtained through Final M-estimator $\hat{\beta}^1$. If an intercept term μ is included in the model, then the RR^2 (MM) is defined as:

$$RR^2(MM) = \frac{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}^0}\right) - \sum \rho\left(\frac{y_i - x_i^T\hat{\beta}}{\hat{s}^0}\right)}{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}^0}\right)} \quad (14)$$

where, $\hat{\mu}$ is the location M-estimate corresponding to the local minimum of:

$$Q_y(\mu) = \sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}^0}\right)$$

such that:

$$Q_y(\mu) < Q_y(\mu^*)$$

where, μ^* is the sample median estimate. RR^2 (MM) also can be obtained from coefficient determination of MM-estimators algorithm. The RVIF which is defined by replacing R^2 in Eq. 12 by RR^2 = (MM) is called the RVIF (MM). Another proposed robust coefficient determination RR^2 = (GM(DRGP)) may be defined as follows:

$$RR^2(GM(DRGP)) = 1 - \frac{\sum_{i=1}^{n} w_{i(GM(DRGP))} r^2_{i(GM(DRGP))}}{\sum_{i=1}^{n} w_{i(GM(DRGP))} (y_i - \bar{y})^2} \quad (15)$$

where,

$$\bar{y} = \sum_{i=1}^{n} w_{i(GM(DRGP))} y_i / \sum_{i=1}^{n} w_{i(GM(DRGP))}$$

r_t and $w_{(GM)}$ are the residual and weight, respectively after the algorithm converged. Subsequently, the RVIF (GM (DRGP)) can be formulated by substituting the R^2 in Eq. 12 with the RR^2(GM(DRGP).

RESULTS

Commercial properties data: In order to investigate the effect of high leverage points on multicolllinearity pattern of the data, a non-collinear data set which was introduced by Kutner *et al.* (2004) is considered. Commercial Properties data containing 81 observations is taken from the suburban commercial properties. The response variable is rental rates which were regressed to the age (X_1), operating expenses and taxes (X_2) and vacancy rates (X_3). This data set contains high leverage points while these high leverage points cannot cause collinearity in this data set (Bagheri and Midi, 2009). According to Bagheri and Midi (2009) adding large magnitude of high leverage points with the same observations to more than two explanatory variables, cause multicollinearity problem for non-collinear data set. Hence, this data set has been modified to have high leverage collinearity-enhancing observations. In order to modify this data set, the first observation of the first two explanatory variables is replaced with a large value of high leverage point (equal to 300). Figure 1a and b present the scatter plot of the original and the modified commercial properties data set.

The coefficient estimations, standard deviations and t-values of the original and the modified commercial properties data set are presented in Table 1 and 2, respectively. The Bootstrap standard deviation of the MM- and our proposed GM (DRGP)-estimates are also computed (Anderson, 2008).

The classical and the Robust VIFs for the original and the modified Commercial Properties data set are exhibited in Table 3.

To verify the merit of our new robust VIF method, a Monte Carlo simulation will be carried.

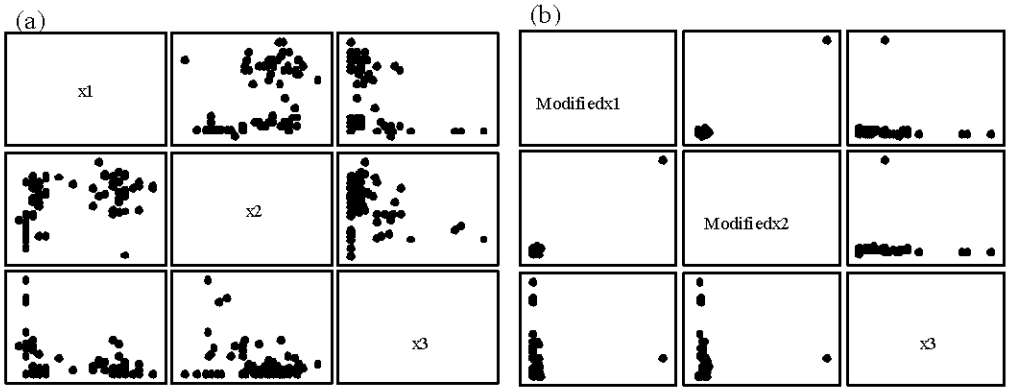

Fig. 1: The scatter plot of (a) original and (b) modified commercial properties data set

Table 1: Parameter estimations, standard deviations and t-values of original commercial properties data set

	Original data set								
	OLS			MM			GM (DRGP)		
Parameters	Estim.	SE	t-value	Estim.	Boot.SE	t-value	Estim.	Boot.SE	t-value
Intercept	11.5407	0.6730	17.1473	11.5550	0. 5752	20.0889	11.3706	0.5243	21.6872
β_1	-0.1190	0.0249	-4.7778	-0.1169	0.0212	-5.5221	-0.1204	0.0353	-3.4108
β_1	0.4461	0.0669	6.6688	0.4137	0.0561	7.3806	0.4471	0.0523	8.5488
β_1	2.6204	1.2229	2.1428	3.9966	1.0268	3.8923	3.7207	1.4225	2.6156

Table 2: Parameter estimation and standard deviation of modified commercial properties data set

	Modified data set								
	OLS			MM			GM (DRGP)		
Parameters	Estim.	SE	t-value	Estim.	Boot.SE	t-value	Estim.	Boot.SE	t-value
Intercept	14.9387	0.2180	68.5339	11.6510	0.6129	19.0104	11.588	0.2466	46.9911
β_1	-0.1268	0.0287	-4.4207	-0.1171	0.0221	-5.2922	-0.1218	0.0203	-6.0000
β_1	0.1236	0.0294	4.2066	0.4056	0.0594	6.8268	0.4298	0.0282	15.2411
β_1	0.2315	1.3022	0.1777	3.9251	1.0708	3.6657	3.5635	1.0727	3.3220

Table 3: Classical and robust VIF for original and modified commercial properties data set

	Original data set			Modified data set		
Variables	CVIF	RVIF (MM)	RVIF(GM (DRGP))	CVIF	RVIF (MM)	RVIF (GM DRGP))
X_1	1.1963	1.2058	1.6345	29.7729	1.1982	1.6006
X_2	1.3086	1.3142	1.9805	29.7844	1.3023	1.9877
X_3	1.1865	1.0643	1.4409	1.0128	1.0643	1.0164

Monte Carlo simulation study: A simulation study is conducted to further asses the performance of our new proposed Robust VIF. Three explanatory variables were considered in which each variable was generated from N (0, 1). We refer to this generated data as the clean independent variables. In order to create collinearity-enhancing observations, the clean data is replaced by certain percentage of high leverage points. The level of high leverage points varied from zero to 25. We considered moderate sample size equals to 100 and 10000 replications in each simulation run. The Magnitude of Contamination (MC) in the X-direction has been varied from 20, 50, 100 and 300. In order to obtain collinearity-enhancing observations, the high leverage points were replaced in all three explanatory variables. The average values of the classical and robust VIF were computed over 10000 simulation runs. To be certain on the result of the simulation study, the percentage of errors for the classical and the robust diagnostics method are calculated to indicate the degree of multicollinearity, whether severe or moderate.

Table 4 shows the performance of our new proposed robust methods that is RVIF (MM) and RVIF (GM (DRGP)) when there aren't any high leverage points in the data set and MC is equal to 100.

Table 5 and 6 exhibit the performance of classical and robust VIF for the small to moderate magnitude of contamination, that is 20 and 50 and moderate to large magnitude of contamination, that is 100 to 300, respectively.

Table 4: The performance of classical and robust VIF methods in the non-collinear data set

Variables	CVIF	RVIF (MM)	RVIF (GM(DRGP))
X_1	1.0572	1.0561	1.0547
X_2	1.0589	1.0206	1.1093
X_3	1.0040	1.0021	1.0057

Table 5: The effect of different percentage and magnitude of high leverage points equal to 20 and 50 on classical and robust VIF when n=100

	MC					
	20			50		
%HL	CVIF	VIF (MM)	VIF (GM (DRGP))	CVIF	VIF (MM)	VIF (GM (DRGP))
5	14.0016	1.0012	1.0349	80.2953	1.0023	1.0501
	14.6837	**1.0141**	**1.0640**	89.8502	**1.0365**	1.0580
	13.7949	1.0105	1.0454	82.5887	1.0072	**1.0634**
	28.5414	1.1077	1.0445	169.4758	1.1625	1.0560
10	27.7636	1.1611	**1.0477**	188.7753	**1.1767**	**1.0925**
	27.4887	**1.2077**	1.0471	177.7806	1.1396	1.0422
15	40.6783	1.3811	1.0797	257.7389	1.3764	1.0431
	49.8620	1.4512	**1.1685**	295.6200	**1.4157**	**1.0718**
	47.6390	**1.4565**	1.0434	254.0743	1.3274	1.0276
20	53.1311	1.4219	1.0266	324.5606	1.5789	1.0654
	46.5985	**1.6345**	1.0413	373.9700	**1.5963**	**1.0867**
	49.5296	1.5569	**1.0936**	342.1948	1.5846	1.0736
25	69.3961	1.6798	1.1397	422.6245	1.7177	1.0532
	73.0520	**2.5565**	**1.3074**	433.6863	1.6496	**1.0655**
	74.7884	1.6754	1.0964	435.6434	**2.3675**	1.0560

#%HL: Percentage of high leverage points, MC: Magnitude of contamination

Table 7 displays the percentage of error for each of the multicollinearity diagnostics method in the identification of the degree of collinearity when the percentage of high leverage point is 25 and different magnitude of high leverage points.

Table 6: The effect of different percentage and magnitude of high leverage points equal to 100 and 300 on classical and robust VIF when n = 100

	MC					
	100			300		
%HL	CVIF	VIF (MM)	VIF (GM (DRGP))	CVIF	VIF (MM)	VIF (GM (DRGP))
5	322.4572	1.0198	1.0559	2998.2889	1.0000	1.0525
	384.1344	**1.0325**	**1.0766**	3525.6178	**1.0288**	**1.0638**
	349.5202	1.0090	1.0479	3020.0975	1.0161	1.0637
	698.2836	1.2196	1.0487	5936.0864	1.1355	1.0268
10	797.7630	**1.2522**	1.0425	5459.1593	1.1771	**1.0414**
	740.3112	1.0740	**1.0608**	5495.6617	**1.1207**	1.0305
15	1034.4164	1.3768	1.0403	9564.2066	1.3869	1.0434
	1078.0602	1.3969	**1.0462**	10090.1482	**1.4158**	**1.0675**
	962.1880	**1.4376**	1.0351	9111.8994	1.4094	1.0275
20	1366.8227	1.5692	1.0417	12052.7027	**1.8469**	**1.0729**
	1509.8460	**1.5699**	**1.0480**	13593.6873	1.6553	1.0679
	1314.4969	1.5235	1.0280	12191.3767	1.7635	1.0662
25	1713.0974	**1.7233**	1.0542	15438.9986	1.6660	1.0474
	1816.7647	1.6953	1.0581	15282.3881	1.6758	**1.0750**
	1715.1178	1.6837	**1.0687**	15555.7704	**1.8091**	1.0382

#%HL: Percentage of high leverage points, MC: Magnitude of contamination

Table 7: The error percentage of classical and Robust VIF for diagnosing collinearity pattern of the data set in 10000 replications for 25 percent high leverage points

		MC			
VIF methods	Degree of muticollinearity	20	50	100	300
CVIF	Non-collinearity (less than 5)	0	0	0	0
	Moderate (between 5 and 10)	0	0	0	0
	Severe (equal or than greater 10)	100	100	100	100
VIF(MM)	Non-collinearity (less than 5)	80.4	85.8	99.999	100
	Moderate (between 5 and 10)	19.6	14.2	0.001	0
	Severe (equal or than greater 10)	0	0	0	0
VIF	Non-collinearity (less than 5)	100	100	100	100
(GM(DRGP))	Moderate (between 5 and 10)	0	0	0	0
	Severe (equal or than greater 10)	0	0	0	0

DISCUSSION

At first we discuss the results of the numerical examples. According to Figure 1a, none of the explanatory variables in the original data set are collinear (Kutner *et al*., 2004). It is interesting to see that as soon as the data is modified, the added high leverage points in x_1 and x_2 pull the regression line toward themselves and change the collinearity pattern of the data which obviously can be seen from Fig. 1b.

It is worth mentioning that for the original data set the F-test is significant (F-value=17.53 and p-value=0.0) which indicates that a linear relationship exist between our variables in the model. From Table 1, that the OLS parameter estimations are significant at significance level of 5% when we compare t-values with t (0.975, 77) =1.99. The results of the coefficient estimations of the two robust methods are also significant. Thus the classical and the robust methods confirm that none of the coefficient estimations are zero. After modifying the data set, the F-test is significant (F-value = 14.75 and p-value = 0.0) while, β_3 for the OLS is not significant (t-value = 0.1777). This behavior is an indicator for the existence of multicollinearity in the data set (Montgomery *et al*., 2001; Kutner *et al*., 2004; Chatterjee and Hadi, 2006). However, the results of Table 2 suggested that both robust methods coefficient determinations are significant. These indicate that the robust methods are fitting the model to the majority of the data and resistant to the high leverage points (Maronna *et al*., 2006; Andersen, 2008). Consequently, these points can't cause any change in the parameter estimates.

The results of Table 3 indicate that when this data set doesn't contain any collinearity-enhancing observations, the classical VIF, RVIF (MM) and RVIF (GM (DRGP)) are not exceeding their cutoff points. Thus these results confirm that this data set is non-collinear data set. However, after the data is modified by creating high leverage points which cause collinearity in the data set, the classical VIF indicates severe multicollinearity (Midi *et al*., 2009; Imon *et al*., 2009). However, our proposed RVIF (GM (DRGP)) and RVIF (MM) are resistant to these added high leverage points and doesn't show collinearity for the data compared to the CVIF which indicates severe multicollinearity. It is evident from the results that the high leverage points are the source of multicollinearity in the data set. The results reveal that the high leverage points which are claimed by Kamruzzaman and Imon (2002) to cause multicollinearity in non-collinear data set, is based on the classical VIF which is not resistant to these points.

Finally we discuss the results obtained from the simulations. According to Table 4, it is important to note that in normal situations, the values of the RVIF (MM) and RVIF (GM (DRGP)) are close to the classical VIF which indicates that they are as good as CVIF in diagnosing the collinearity pattern of the data correctly. As to be expected, when there aren't any high leverage points in the data set, the CVIF confirmed that the data set is not collinear (Table 4) (Hair *et al*., 2006). It can be observed from Table 5 and 6 that by increasing the percentage and magnitude of high leverage points, the value of CVIF become larger. Hence, the CVIF is very sensitive to the presence of added high leverage points to the data set. However, the RVIF (MM) increases for the small values of high leverage points such as 20 while for the large values of high leverage points such as 300 it doesn't change drastically (Table 6). Hence, the RVIF (MM) for small values of high leverage points is not as resistant as high values.

It is evident from the results that the RVIF (GM (DRGP)) is not affected by the increased in the percentage and magnitude of high leverage points.

It is important to note that a method that has very small or high percentage of error reveals that it can detect correctly or unable to detect correctly the degree of multicollinearity in the data, respectively. Inevitably, robust methods have high percentage of error due to their resistance to the high leverage points. It can be observed from Table 7 that the percentage of errors for the CVIF is zero for different magnitude of high leverage points considered in this study. The results reveal that the CVIF has detected a severe multicollinearity due to their sensitivity to the added high leverage points in the data sets. On the other hands, the RVIF (MM) has detected a moderate degree of multicollinearity for small magnitude of high leverage point. This shows that this method is not robust against small magnitude of high leverage points. However, by increasing the magnitude of high leverage points up to 300, it becomes resistant to these points. The results also point out that the RVIF (GM (DRGP)) always robust against high leverage collinearity enhancing observations.

CONCLUSIONS

In this study, we proposed a robust RVIF (MM) and RVIF (GM (DRGP)) for detecting the source of multicollinearity which is caused by the high leverage points in the data set. Recently, high leverage points are known to be another source of multicollinearity. The results of the study signify that the high leverage points has an unduly effects on the classical multicollinearity diagnostics, specifically the VIF. In this situation, sometimes the classical VIF conveys the misleading interpretation about collinearity pattern of the data. The results of the real data and simulation study reveal that the high leverage points are the source of multicolinearity evidently by the failure of the RVIF in diagnosing multicollinearity, contradicts to the result of the CVIF which indicates existence of multicollinearity in the data set. Another important conclusion is that the RVIF (GM (DRGP)) is the most resistant diagnostic measure to high leverage points, followed by the RVIF (MM).

REFERENCES

Andersen, R., 2008. Modern Methods for Robust Regression. SAGE Publications, USA., ISBN: 9781412940726.

Bagheri, A. and H. Midi, 2009. Robust estimations as a remedy for multicollinearity caused by multiple high leverage points. J. Math. Statist., 5: 311-321.

Belsley, D.A., E. Kuh and R.E. Welsch, 1980. Regression Diagnostics: Identifying Influential Data and Sources of Colinearity. John Willey and Sons Inc., New York.

Chatterjee, S. and A.S. Hadi, 2006. Regression Analysis by Example. 4th Edn., Wiley, New York.

Coakley, C.W. and T.P. Hettmansperger, 1993. A bounded-influence, high-breakdown, efficient regression estimator. J. Am. Statist. Assoc., 88: 872-880.

Croux, C., P.J. Rousseeuw and O. Hossjer, 1994. Generalized S-estimators. J. Am. Statist. Assoc., 89: 1271-1281.

Hadi, A.S., 1988. Diagnosing collineariy-influential observations. Comput. Statist. Data Anal., 7: 143-159.

Hair, J.F., R. Anderson, R.L. Tatham and W.C. Black, 2006. Multivariate Data Analysis. Prentice Hall, Upper Saddle River, ISBN: 10: 0138948585.

Hill, R.W., 1977. Robust regression when there are outliers in the carriers. Ph.D. Thesis, Harvard University, Boston, MA.

Imon, A.H.M.R., A. Bagheri and H. Midi, 2009. Two-step robust diagnostic method for identification of multiple high leverage points. J. Math. Statist., 5: 97-106.

Jolliffe, I.T., 1982. A note on the use of principal components in regression. J. R. Statist. Soc. Ser. C, 31: 300-303.

Kamruzzaman, M. and A.H.M.R. Imon, 2002. High leverage point: Another source of multicollinearity. Pak. J. Statist., 18: 435-448.

Kutner, M.H., C.J. Nachtsheim and J. Neter, 2004. Applied Linear Regression Models. 4th Edn., McGraw Hill, New York, ISBN: 978-0256086010.

Maronna, R.A., R.D. Martin and V.J. Yohai, 2006. Robust Statistics Theory and Methods. Willy and Sons, New York, ISBN: 10: 0470010924.

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

Midi, H., 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 Statist., 36: 507-520.

Montgomery, D., E. Peck and G.G. Vining, 2001. Introduction to Linear Regression Analysis. 3rd Edn., Jon Wiley and Sons, New York.

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

Rousseeuw, P.J. and V.J. Yohai, 1984. Robust Regression by Means of S-Estimators. In: Robust and Nonlinear Time Series Analysis, Franke, J., W. Hurdle and R.D. Martin (Eds.). Springer-Verlag, New York, USA., pp: 256-272.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. Math. Statist. Appl., 13: 283-297.

Rousseeuw, P. and B. van Zomeren, 1990. Unmasking multivariate outliers and leverage points. J. Am. Stat. Assoc., 85: 633-639.

Rousseeuw, P.J. and M. Hubert, 1997. Recent Development in PROGRESS. In: L1-Statistical Procedures and Related Topics, Dodge, Y. (Ed.). Vol. 31, Institute of Mathematical Statistics, Hayward, California, pp: 201-214.

Rousseeuw, P.J. and A.M. Leroy, 2003. Robust Regression and Outlier Detection. John Willy, New York, ISBN-10: 0471852333.

Simpson, J.R., 1995. New methods and comparative evaluations for robust and biased-robust regression estimation. Ph.D. Thesis, Arizona State University.

Wilcox, R.R., 2005. Introduction to Robust Estimation and Hypothesis Testing. 2nd Edn., Elsevier Academic Press, USA., ISBN: 0-12-751542-9.

Yohai, V.J., 1987. High breakdown-point and high efficiency robust estimates for regression. Ann. Statist., 15: 642-656.