Sains Malaysiana 40(12)(2011): 1437–1447

A Monte Carlo Simulation Study on High Leverage Collinearity-Enhancing Observation and its Effect on Multicollinearity Pattern

(Kajian Simulasi Monte Carlo Terhadap Cerapan Titik Tuasan Tinggi yang Mempertingkatkan Kolinearan dan Kesannya Terhadap Pola Multikolinearan)

HABSHAH MIDI, AREZOO BAGHERI*
& A.H.M. RAHMATULLAH IMON

ABSTRACT

Outliers in the X-direction or high leverage points are the latest known source of multicollinearity. Multicollinearity is a nonorthogonality of two or more explanatory variables in multiple regression models, which may have important influential impacts on interpreting a fitted regression model. In this paper, we performed Monte Carlo simulation studies to achieve two main objectives. The first objective was to study the effect of certain magnitude and percentage of high leverage points, which are two important issues in tending the high leverage points to be collinearity-enhancing observations, on the multicollinarity pattern of the data. The second objective was to investigate in which situations these points do make different degrees of multicollinearity, such as moderate or severe. According to the simulation results, high leverage points should be in large magnitude for at least two explanatory variables to guarantee that they are the cause of multicollinearity problems. We also proposed some practical Lower Bound (LB) and Upper Bound (UB) for High Leverage Collinearity Influential Measure (HLCIM) which is an essential measure in detecting the degree of multicollinearity. A well-known example is used to confirm the simulation results.

Keywords: Collinearity influential measure; collinearity influential observations; condition number; diagnostic Robust Generalized Potential (DRGP) method; high leverage points

ABSTRAK

Titik terpencil arah X atau titik tuasan tinggi adalah punca terkini bagi multikolinearan. Multikolinearan berlaku apabila dua atau lebih pembolehubah tak bersandaran dalam model regresi berganda tak berortogonal, yang mungkin memberi pengaruh penting ke atas interpretasi model regresi tersuai. Dalam kertas ini, kami menjalankan kajian simulasi Monte Carlo untuk mencapai dua objektif utama. Objektif pertama ialah untuk mengkaji kesan magnitud tertentu dan peratus titik tuasan tinggi ke atas pola data, yang mana keduanya adalah dua isu penting yang menjuruskan titik tuasan tinggi kepada cerapan yang mempertingkatkan kolinearan. Objektif kedua adalah untuk mengkaji situasi bagaimana titik tuasan ini menjadikan tahap multikolinearan berbeza, seperti sederhana atau tinggi. Berpandukan kepada keputusan simulasi, titik tuasan tinggi sepatutnya mempunyai magnitud yang besar bagi sekurang-kurannya dua pembolehubah takbersandaran untuk memastikan mereka adalah penyebab masalah multikolinearan. Kami juga mencadangkan Batas Bawah (LB) and Batas Atas (UB) bagi Ukuran Titik Tuasan Tinggi Berpengaruh Kolinearan (HLCIM) yang menjadi ukuran penting untuk mengesan tahap multikolinearan. Contoh terkenal digunakan untuk menentusahkan keputusan simulasi.

Kata kunci: Cerapan yang mempertingkatkan kolinearan; kaedah Potensi Teritlak Teguh Berdaignostik (DRGP); nombor kondisi; titik tuasan tinggi; ukuran kolinearan berpengaruh

INTRODUCTION

Multicollinearity or nonorthogonality is a near-linear dependency between two or more explanatory variables. Its presence causes difficulties in making prediction inferences and estimations as well as selecting an appropriate set of variables for the model. Unfortunately, in most regression applications the explanatory variables are not orthogonal. In such cases, any inferences based on the parameter estimations of the model become invalid. There are several sources of multicollinearity. Montgomery et al. (2001) noted that multicollinearity may be due to the data collection method employed, constraints on the model or in the population being sampled, model specification such as adding polynomial terms to the regression model, and an over determined model which is defined as a model with more explanatory variables than the number of observations. It is important to note that multicollinearity is a problem which exists in the data set, thus, there is no statistical test for its presence. However, a diagnostic method can replace a statistical test to indicate the existence and extent of multicollinearity in the data set. A very simple measure of multicollineariy is the examination of the correlation matrix of explanatory variables. Although, when more than two explanatory variables are involved in a

near-linear dependence, there is no assurance that any of the pairwise correlation coefficients will be large (Montgomery et al. 2001). It is worth mentioning that collinearity may exist even if all the pairwise correlations are insignificant. However, the presence of pairwise correlations may be a significant sign for the existence of multicollinearity. Marquardt (1970) proposed Variance Inflation Factor (VIF) as another popular diagnostic tool of multicollinearity. The VIF measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. Moreover, a practical and useful multicollinearity diagnostic method such as Condition Number (CN) of X matrix could be obtained from the eigen structure analysis of cross-products X matrix. Therefore, $X'X$ matrix may be factored into p ordered eigenvectors and eigenvalues. The first eigenvector is a linear combination of the independent variables that explains the possible maximum variance. This eigenvector is associated with the largest eigenvalue. Subsequent eigenvectors maximize the remaining variance and are associated with smaller eigenvalues. An eigenvalue of zero indicates a perfect multicollinearity. Belsley et al. (1980) proposed a similar approach for diagnosing multicollinearity. The singular-value decomposition could be useful in identifying CN of X matrix. Belsley (1991) performed some experiments to discover whether the diagnostic methods could identify multicollinearity (or not) and which variables were also involved in the multicollinearity. He aimed to provide guidance on how high the condition number should be to indicate a multicollinearity problem in the data set.

Kamruzzaman and Imon (2002) introduced a new source of multicollinearity, which is high leverage points, observations not only deviated from the same regression line as the other data but also fall far from the majority of explanatory variables in the data set (Hocking & Pendelton 1983; Moller et al. 2005). They studied this new source of multicollinearity through simulated and real data sets. Utilizing the correlation matrix, they proved that the presence of multiple equal or unequal high leverage points causes severe multicollinearity. Unequal high leverage points may cause more multicollinearity problems than the equal high leverage cases.

According to Hadi (1988), these new sources of multicollinearity may be collinearity influential observations. Hadi (1988) noted that the collinearity influential observations are usually points with high leverages while all high leverage points are not collinearity influential observations. Sengupta and Behimasankaram (1997) pointed out that the weakness of this measure is in the lack of symmetry, which is due to the additive change in CN of X matrix.

Yet, little attention has been devoted to the role of the individual cases in collinearity of explanatory variables in the data set (Sengupta & Bhimasankaram 1997). Furthermore, there is a lack of investigation in the literature on high leverage points that cause multicollinearity problems. Hence, two important issues were investigated in this study. The first issue was to investigate in which conditions the high leverage points tend to become collinearity influential, specifically to be collinearity-enhancing observations. The second issue was to examine the effect of the high leverage collinearity influential observation, which is a new source of multicollinearity, on the most applicable multicollinearity diagnostics such as CN of X matrix. In this way, we can investigate the degree of multicollinearity caused by the high leverage points. Unfortunately, there is no direct technique to determine in which situations high leverage points may cause multicollinearity and also how we can investigate the degrees of multicollinearity caused by these points. Insight is gained only by simulation experiences and by real data sets.

Materials and Methods

High Leverage Collinearity Influential Measure

Regression model can be defined as:

$$Y = X\beta + \varepsilon \tag{1}$$

where Y is an ($n\times1$) vector of response or dependent variables, X is an ($n\times p$) ($n>p$) matrix of predictors (collinear explanatory variables), β is a ($p\times1$) vector of unknown finite parameters to be estimated and ε is an ($n\times1$) vector of random errors. We let the j^{th} column of the X matrix be denoted as X_j, therefore $X = [X_1, X_2,\ldots,X_p]$. Additionally, we defined multicollinearity in terms of the linear dependence of the columns of X, i.e., whereby the vectors of X_1, X_2 ,…,X_p are linearly dependent if there is a set of constants t_1, t_2,…,t_p , that are not all zero, such as:

$$\sum_{j=1}^{p} t_j X_j = 0. \tag{2}$$

If (2) holds exactly, we face severe multicollinearity problem. However, the problem of moderate multicollinearity is said to exist when (2) holds approximately.

Marquardt (1970) proposed the Variance Inflation Factor (VIF) as multicollinearity diagnostic tool which is defined as follows:

$$VIF_j = \frac{1}{1-R_j^2} \qquad j = 1, \ldots, k. \tag{3}$$

Where R^2 is the coefficient determination of each of the explanatory variables when regressed on the other explanatory variables by using the ordinary least squares method. Any VIF value between 5 and 10 indicates that moderate multicollineriaty exists in the data set. Severe multicollinearity may happen when VIF exceeds its cutoff point 10.

Condition Number (CN) of $X'X$ matrix which is another useful multicollinearity diagnostic method could be obtained as follows. If eigenvalues are formed into Condition Indices (CI) for $j = 1$ to p and we also introduce

the eigenvalues of matrix $X'X$ as $\lambda_1, \lambda_2, \ldots, \lambda p$ then the CI of matrix $X'X$ is:

$$k_j = \frac{\lambda_{\max}}{\lambda_j} \qquad j = 1, \ldots, p \tag{4}$$

k is the largest CI, which is known as the CN of $X'X$ matrix. To make the condition indices comparable from one data set to another, the independent variables should first be scaled by dividing each of the explanatory variables with its standard deviation, to have the same length. Scaling will prevent the eigen analysis to be dependent on the variables units of measurements. Subsequently, they may also be centered by correcting X for its average $\bar{X}$. Nonetheless, the choice of centering is somehow arbitrary, since some authors argued that centering removes any collinearity that involves the intercept. By centering, the intercept will be removed from the regression model and consequently removing any collinearity which may exist between intercept and the other explanatory variables (Belsley 1984; Montgomery et al. 2001).

Belsley et al. (1980) identified the singular-value decomposition of $(n \times p)$ X matrix as:

$$X = UDV', \tag{5}$$

where U (the matrix which columns are the eigenvectors associated with the p non-zero eigenvalues of $X'X$) is $(n \times p)$, V (the matrix of eigenvectors of $X'X$) is $(p \times p)$, $U'U = I$, $V'V = I$, and D is a $(p \times p)$ diagonal matrix with non-negative diagonal elements $\mu_j, j = 1, 2, \ldots, p$ which is called *singular-values* of X. They also defined the CI of the X matrix as:

$$\eta_j = \frac{\mu_{\max}}{\mu_j} = \sqrt{k_j} \qquad j = 1, \ldots, p, \tag{6}$$

where $\eta_1, \eta_2, \ldots, \eta_p$ are the singular values of X matrix. It is noticeable that the largest value of k_j and also η_j can be defined as CN of matrix $X'X$ and X matrix, respectively.

Belsley (1991) recommended that CN between 10 and 30 for X matrix be indicated as moderate multicollinearity while more than 30 results as severe multicollinearity. This was the first attempt to give meaning to the value of multicollinearity diagnostic. The author's rule of thumb has been accepted as the standard in application. However, there were several limitations to the experiments such as the small number of experiments, which varied only by degree of multicollinearity and sample size. Many studies have been devoted to this issue (Mason & Perreault 1991; Rosen 1999; Schindler 1986; Stinnett 1993). It is worth mentioning that CN has been used in this article as CN of X matrix.

Hadi (1988) defined a measure for the influence of the i^{th} row of X matrix on the condition index as:

$$\delta_i = \frac{k_{(i)} - k}{k} \qquad i = 1, \ldots, n. \tag{7}$$

where $k_{(i)}$ can be computed from the eigenvalues of $X_{(i)}$ when the i^{th} row of X matrix has been deleted. Hadi (1988) reported that a large negative value of δ_i indicates that group i is a collinearity-enhancing observation while a large positive δ_i value indicates a collinearity-reducing set. Nevertheless, there was no mention specifically as to how large the values of δ_i should be. In this respect the usage of Hadi's measure is not practical because its cutoff-points are based on the researcher's judgment on the magnitude of the δ_i.

To overcome the lack of symmetry problem of Hadi's measure, Sengupta and Behimasankaram (1997) proposed a new collinearity influential measure as:

$$l_i = \log\left(\frac{k_{(i)}}{k}\right) \qquad i = 1, \ldots, n. \tag{8}$$

Although they didn't propose any specific cutoff point for l_i, they introduced some easily computable lower bound and upper bound values for this new collinearity influential measure (Sengupta and Behimasankaram 1997) . Following the idea of Sengupta & Bhimasankaram (1997), we proposed a new measure which is called the High Leverage Collinearity Influential Measure (HLCIM) and is defined as follows:

$$\text{HLCIM} = \log\left(\frac{k_D}{k}\right), \tag{9}$$

where D is the group of high leverage Collinearity Influential Observations. The HLCIM is used as an indicator to indicate whether high leverage points can cause multicollinearity or not in the data set. Monte Carlo simulations may be a good approach to define the cutoff point for the HLCIM. It is important to note that high leverage points can hide and induce multicollinearity pattern in two different situations. The first situation is when $\frac{k_{(D)}}{k} > 1$ and $k_{(D)} > k$, then $\log\left(\frac{k_{(D)}}{k}\right) > 0$ which result in the deletion of the high leverage points. Consequently, the degree of multicollinearity increases due to the characteristics of high leverages which hide the multicollinearity pattern. In this situation, the high leverage points are referred as collinearity-reducing observation. Otherwise, the deletion of high leverage points may reduce the degree of multicollinearity. Thus, in this situation the high leverage points are referred as collinearity-enhancing observation and satisfy this inequality; $0 < \frac{k_{(D)}}{k} < 1$ and $k_{(D)}$ then $\log\left(\frac{k_{(D)}}{k}\right) < 0$.

HIGH LEVERAGE DIAGNOSTICS METHODS

A traditional measure of the outlyingness of an observation X_i with respect to the sample is the Three-Sigma edit rule, which is defined as follows:

$$T = \frac{X - \bar{X}}{s}, \tag{10}$$

where $\bar{X}$ is the mean and s is the standard deviation of explanatory variables. The robust version of (10) is:

$$T' = \frac{X - Med(X)}{Mad(X)}, \tag{11}$$

where *Med(X)* is *Median(X)* and *Mad(X)* = 1.4826 (Median $|X_i$ – median $(x_i)|$ is the normalized median absolute deviation about the *Median(X)*. *T* and *T*' are approximately equal, when the distribution of the data is normal. The observation which has absolute value of *T* or *T*' more than 3 is considered as an outlier (Maronna 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 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 the 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 the twice-the-mean-rule ($2k/n$) by Hoaglin &Welsch (1978). However, Hadi (1992) explained that the hat matrix might fail to identify the high leverage points due to the effect of high leverage points in the leverage structure. So, he introduced another diagnostic tool as follows:

$$p_{ii} = \frac{w_{ii}}{1 - w_{ii}}, \tag{12}$$

where $w_{ii} = x_i^T \left(X^T X\right)^{-1} x_i$ is the diagonal element of *W* and the i^{th}, diagonal potential p_{ii} can be defined as $p_{ii} = x_i^T \left(X_{(i)}^T X_{(i)}\right)^{-1} x_i$ where $X_{(i)}$ is the data matrix *X* without the i^{th} row. He proposed a cutoff 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. Still, this method was unable to detect all of the high leverage points.

Imon (2002) introduced another diagnostic tool as generalized potentials for the whole data set, which is:

$$p_{ii}^* = \begin{cases} w_{ii}^{(-D)} & \text{for } i \in D \\ \dfrac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} & \text{for } i \in R \end{cases}, \tag{13}$$

where *D* is a deleted set, meaning any observations which is suspected as outliers and R is the remaining set from observations after deleting $d < (n-p)$ therefore containing $(n-d)$ cases. Because there isn't any finite upper bound for p_{ii}^*'s and the theoretical distribution of them are not easily found, he used a *Mad*-cutoff point for the generalized potential as well.

Recently, Habshah et al. (2009) developed a Diagnostic Robust Generalized Potential (DRGP) to determine outlying points in multivariate data set by utilizing the Robust Mahalanobis Distance (RMD) based on Minimum Volume Ellipsoid (MVE). We refer this method as the DRGP (MVE). The set *D* (deletion set) in generalized potentials method in (13) is defined based on the points which RMD-MVE exceeds Median (RMD-MVE) +3*Mad* (RMD-MVE). Rousseeuw (1985) introduced RMD-MVE as:

$$RMD_i = \sqrt{\left(X - T_R(X)\right)^T C_R(X)^{-1} \left(X - T_R(X)\right)} \; for \; i = 1, \ldots, n, \tag{14}$$

where $T_R(X)$ and $C_R(X)$ are robust locations and shape estimates of the MVE. Then, generalized potential statistics with the Mad-cutoff point has been utilized to check whether all members of the deletion set have potentially high leverage or not. The merit of this method is in swamping less low leverages as high leverage points in the data set. Hence, this method has been utilized in the following chapter as a diagnostic method to define high leverage points.

Results and Discussion

Before proceeding to the simulation study, the effect of high leverage points in multicollinearity pattern of the data will be investigated.

The Effect of High Leverage Points on Multicollinearity

In order to explore the effect of high leverage points on multicolllinearity pattern of the data, a non-collinear data set which was introduced by Neter et al. (2004) is considered. Commercial Properties data containing 81 observations was taken from the suburban commercial properties. The response variable was rental rates which were regressed to the age (X_1), operating expenses and taxes (X_2) and vacancy rates (X_3). This data set contained 19 high leverage points (observations 1, 2, 3, 6, 7, 8, 17, 21, 26, 29, 37, 45, 53, 54, 58, 61, 62, 72 and 79). As already been mentioned, Hadi (1988) pointed out that, influential points are usually the points with high leverages. It was also noted that not all high leverage points are collinearity influential observations and vice- versa. We will show in Figure 1 and Table 1 that these high leverage points are not collinearity influential observations. Hence, the data was modified in three situations; to investigate the effect of adding one high leverage to one explanatory, one high leverage to each of two explanatory and one high leverage to each of three explanatory variables, in inducing collinearity. To create high leverage collinearity-enhancing observations, the data is modified accordingly such that the first observation of each explanatory variable is replaced with 300 for each of the three situations.

Figure 1(a), (b), (c) and (d) display the scatter plot matrix of the original data, modified by a new high leverage in X_1, modified by a new high leverage in X_1 and X_2, modified by a new high leverage in X_1, X_2 and X_3 for the Commercial Properties data set. Let us first focus our attention to the original data in Figure 1(a). We can see from this figure that there isn't any collinearity between explanatory variables. Figure 1(b) illustrates the effect of

this new added high leverage points in X_1. It is interesting to point out that this high leverage point couldn't make collinearity between explanatory variables. Figure 1(c) and Figure 1(d) suggest that the explanatory variables have become collinear.

The collinearity diagnostic methods such as correlation matrix, variance inflation factor and condition indices of the normalized explanatory variables (transformed the explanatory variables to Z-scores) for the original and the modified data sets are presented in Table 1. The result of Table 1 signifies that, although the original data set has 19 high leverage points but these leverage points did not cause multicollinearity problem. This result was supported by Figure 1(a) where the explanatory variables were not correlated. It can be observed from Table 1 that for modified data whereby one high leverage point is added to X_1, none of the diagnostic methods reveal collinearity. On the other hand, when high leverages are added to two and all three explanatory variables, these high leverages not only induce multicollinearity but they also change the degree of multicollinearity from moderate to strong. Thus, the new high leverage points are collinearity-enhancing observations. Furthermore, when we increase the number of explanatory variables with high leverage collinearity-enhancing observations, they may enhance the degree of collinearity between explanatory variables as well.

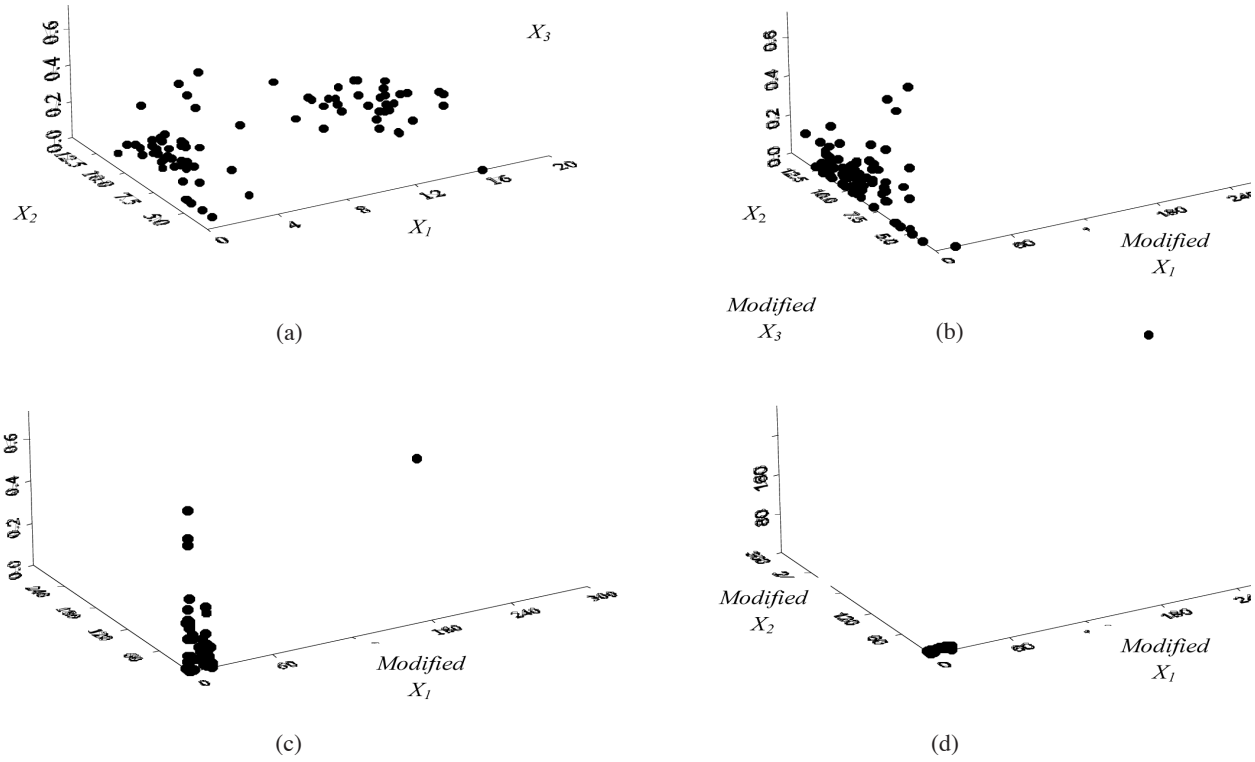

FIGURE 1. Scatter plot matrix of original (a), modified by a new high leverage in X_1 (b) modified by a new high leverage in X_1 and X_2 (c), modified by a new high leverage in X_1 , X_2 and X_3 (d) Commercial Properties data set

TABLE 1. Collinearity diagnostics of the original and modified commercial properties data set

Diagnostics	Status	1	2	3
Pearson correlation coefficient	original data	r_{12}= 0.39	r_{13}= -0.25	r_{23}= -0.38
	New High leverage in X_1	r_{12}= -0.13	r_{13}= -0.00	r_{23}= -0.38
	New High leverage in X_1 and X_2	r_{12}= 0.98	r_{13}= -0.00	r_{23}= 0.02
	New High leverage in X_1 , X_2 and X_3	r_{12}=0.98	r_{13}= 0.98	r_{23}=1.00
VIF > 5	original data	1.20	1.31	1.19
	New High leverage in X_1	1.02	1.19	1.17
	New High leverage in X_1 and X_2	29.77	29.78	1.01
	New High leverage in X_1 , X_2 and X_3	29.41	184.71	157.07
Condition index > 10	original data	1	1.50	1.72
	New High leverage in X_1	1	1.18	1.53
	New High leverage in X_1 and X_2	1	1.41	10.82
	New High leverage in X_1 , X_2 and X_3	1	11.14	31.28

MONTE CARLO SIMULATION STUDY

A Monte Carlo simulation study is designed to achieve four objectives. Two different simulation designs were performed. In both simulation, the first 100(1-α) percent observations of each explanatory variables were generated from a standard normal distribution. We refer to this generated data as the clean explanatory variables. The remaining 100 α percent of the observations were contaminated explanatory variables. To generate contaminated explanatory variables, a Robust Distance (RD) for each of the clean explanatory variable was computed in the first step. In this paper, we call $|Median(x) + 3Mad(x)|$ as Robust Distance (RD), which is approximately equal to 3 for standard normal distribution. For the second step, the RD value was multiplied with different multipliers which are called magnitude of contamination (MC) to produce high leverage points. It is worth mentioning that equal MCs for different explanatory variables have been considered to ease the computation. In each simulation run, there were 10,000 replications. A real well- referred data set is applied to verify the simulation results.

High leverage collinearity influential observations in multiple linear model with three explanatory variables and n=100 A model with three explanatory variables with a moderate sample size of 100 and the MC values from 1 to 4 were considered. The percentage of contamination in each explanatory variable was 5, 10, 15, 20 and 25 percent.

The first objective of this simulation study is to determine at which magnitude and percentage of contamination will the contamination points be detected as high leverage points in situations where contamination exists in one, two and all three explanatory variables.

Table 2 illustrates the percentage of multiple high leverage points detected by the DRGP (MVE) at different magnitudes and percentage of contaminations. The results based on contaminated explanatory variable X_1, variables X_1 and X_2, and X_1, X_2 and X_3 are shown on the Table 2. It can be observed from Table 2 that when the magnitude of contamination (MC) is 3 or more, and contamination exists in one, two or three explanatory variables, the DRGP (MVE) detects these points as high leverage points irrespective to the percentage of high leverage points. It is worth mentioning here that by increasing the number of contaminated explanatory variables in the model, contaminated observations become high leverage points in a smaller value of MC. For instance, in Table 2, when we had one contaminated explanatory variable in the model, it can be seen that the MC value equal to almost 3, making the contaminated points multiple high leverage points. Subsequently, MC equal to 2 is enough to make two and more contaminated explanatory variables to be multiple high leverage points. Hence, it is obvious that any points with large MC in any explanatory variables are detected as high leverage points. Whilst, it is noticeable that any high leverage points detected by DRGP (MVE) should not necessarily have large MC.

The second objective is to determine whether the high leverage points which exist in one or all explanatory variables were collinearity-enhancing observations. Table 3 presents the effect of different levels of magnitude and different percentage of contamination on HLCIM and CN in three explanatory variables model, where contamination is in one explanatory variable. The MC value displayed in Table 3 starts from 3 onwards and does not include MC less than 3 because the result of Table 2 had suggested that high leverage points in one explanatory variable are correctly identified for value of MC equal and greater than 3. Here, we wanted to investigate by means of HLCIM and condition number CN of X matrix whether high leverage points in one explanatory variable were collinearity-enhancing observation. The small and positive values of all HLCIM in Table 3 show that when high leverage points are in one explanatory variable, these points may not cause multicollinearity. This result is in agreement with the small value of condition numbers (more than the 10 cutoff point for moderate multicollinearity), which suggests that the high leverage points are not collinearity-enhancing observations. Moreover, in this situation, if the percentage and magnitudes of high leverage points increases, the HLCIM doesn't change drastically. Thus, it is obvious that the high leverage points can't be collinearity-enhancing observations when high leverage points exist in one explanatory variable in three explanatory variables model.

TABLE 2. The percentage of multiple high leverage points detected by DRGP (MVE) in three explanatory variables model, n=100

CEV	α	MC				CEV	α	MC				CEV	α	MC			
		1	2	3	4			1	2	3	4			1	2	3	4
		DRGP(MVE)						DRGP(MVE)						DRGP(MVE)			
	5	2.78	5	5	5		5	4.74	5	5	5		5	5	5	5	5
	10	3.48	10	10	10	X_1	10	8.3	10	10	10	X_1	10	10	10	10	10
X_1	15	1.4	15	15	15	X_2	15	18.72	15	15	15	X_2	15	14.78	15	15	15
	20	2.3	19.77	20	20		20	5.8	20	20	20	X_3	20	19.67	20	20	20
	25	2.31	24.91	25	25		25	2.75	25	25	25		25	24.87	25	25	25

CEV# Contaminated Explanatory Variable, MC# Magnitude of Contamination

The third objective is to study the effect of MC and percentage of high leverage points on HLCIM and CN and subsequently propose cutoff points for the HLCIM in the certain number of sample size (100) and number of explanatory variables (3) so that it can be used as an indicator of whether the degrees of multicollinearity caused by collinearity-influential observations were moderate or severe.

Table 4 exemplifies the HLCIM and CN of three variables model when high leverage points are in all three explanatory variables (X_1, X_2 and X_3). Here, we also wanted to investigate the effect of different magnitudes and different percentage of contamination on the value of HLCIM and CN when contamination is in all explanatory variables in three explanatory models with a sample size equals to 100. Following the results of Table 2, only values of MC with magnitude of two are presented in Table 4, as this contamination points are high leverage points. It can be observed from Table 4, when high leverage points exist in all three explanatory variables, by increasing the percentage of high leverage points and magnitudes of MC, the value of $\log\left(\frac{k_{(D)}}{k}\right)$, or HLCIM starts from negative and become larger. In this situation, the value of CN also becomes larger.

From these results we propose to define the Lower Bound (LB) value of the HLCIM when the corresponding CN becomes 10. This LB can be used as an indicator of moderate multicollinearity. The Upper Bound (UB) value of the HLCIM is defined when the corresponding CN becomes 30 and it can be used as an indicator of severe multicollinearity. Computing the average of the all LB and UB values of HLCIM for different percentage of high leverage points, result in the values of -0.90 and -1.34 as LB and UB values of HLCIM respectively. Thus, in three contaminated explanatory variables model with MC more than 2 and a sample size equaling to 100, negative value of HLCIM which is more than 0.94 indicates moderate multicollinearity, and when it is negative and more than 1.34 indicates that severe multicollinearity exists in the data set. In this situation the high leverage points will become high leverage collinearity-enhancing observations as well. It can be seen that by increasing the percentage of high leverage points, the high leverage collinearity-enhancing observations can be detected at a smaller MC, and vice-versa. For example, when all three explanatory variables have 5 percent high leverage with the MC more than 8, these high leverages can cause multicollinearity, while the high leverage points will be the high leverage collinearity-enhancing in 15 percent high leverage with a magnitude of contamination more than 5.

Figure 2 presents the effect of MC on CN. By looking at Table 4 and Figure 2, 5 percent of high leverage points with an average magnitude more than 8 in three contaminated explanatory variables will be collinearity-enhancing observations,which bring moderate multicollinearity problems, and if the magnitude of contamination is more than 24, these high leverage collinearity enhancing observations will bring severe multicollinearity in the data set. Similar results can be drawn for different percentage of high leverage points. The results also pointed out that the CN values become 10 at the very least which brings to moderate multicollinearity and make these high leverage points collinearity-enhancing observations for 10, 15, 20 and 25 percentage of high leverage points corresponding to MC values which are more than 6, 5, 4 and between 3 and 4, respectively. Accordingly, the values of MC approximately more than 17, 14, 12, and 11 for 10, 15, 20 and 25 percent high leverage points can cause severe multicollinearity.

The findings are summarized as follows. In the situation with three contaminated explanatory variables model, sample size equals to 100 and the contaminated points with MC more than 2, these points are detected as high leverage points. If the HLCIM is negative and more than 0.90 (LB), these points are detected as high leverage collinearity-enhancing observations which make moderate multicollinearity. On the other hand, the negative value of HLCIM which is more than 1.34 indicates the existence of severe multicollinearity between explanatory variables.

High leverage collinearity- enhancing observations in multiple linear model with more than three explanatory variables and various sample size The objective of this simulation study is to propose cutoff points for HLCIM to

TABLE 3. HLCIM and CN of three explanatory variables model when contamination is in one explanatory variable (X_1), n=100

α	(MC)													
	3		4		5		6		7		8		9	
	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN
5	0.081	1.17	0.081	1.18	0.082	1.18	0.08	1.17	0.078	1.17	0.08	1.18	0.082	1.17
10	0.079	1.17	0.081	1.17	0.080	1.170	0.083	1.17	0.082	1.17	0.083	1.17	0.079	1.18
15	0.082	1.18	0.08	1.17	0.080	1.170	0.078	1.17	0.081	1.17	0.082	1.18	0.08	1.18
20	0.081	1.17	0.081	1.17	0.083	1.18	0.083	1.17	0.081	1.18	0.081	1.17	0.082	1.17
25	0.08	1.18	0.082	1.17	0.083	1.18	0.08	1.17	0.081	1.17	0.078	1.18	0.08	1.17

HLCIM # High Leverage Collinearity- Influential Measure ($\log\left(\frac{k_{(D)}}{k}\right)$) CN# Condition Number of X matrix, MC# Magnitude of Contamination

TABLE 4. HLCIM and CN of three explanatory variables model when contamination is in all three explanatory variable (X_1, X_2 , X_3), n=100

	MC in three explanatory variables															
	(2,2,2)		(3,3,3)		(4,4,4)		(5,5,5)		(6,6,6)		(7,7,7)		(8,8,8)		(9,9,9)	
α	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN
5	-0.28	2.7	-0.44	3.88	-0.55	5.120	-0.65	6.32	-0.73	7.57	-0.79	8.82	-0.90	10.09	-0.93	11.31
10	-0.42	3.7	-0.58	5.43	-0.70	7.16	-0.80	8.91	-0.92	10.78	-0.95	12.51	-1.00	14.22	-1.05	15.96
15	-0.50	4.47	-0.67	6.63	-0.79	8.81	-0.89	11.03	-0.97	13.14	-1.04	15.31	-1.09	17.47	-1.14	19.69
20	-0.56	5.15	-0.73	7.68	-0.90	10.15	-0.95	12.71	-1.03	15.23	-1.10	17.82	-1.16	20.26	-1.21	22.83
25	-0.61	5.78	-0.78	8.57	-0.91	11.42	-1.00	14.22	-1.08	17.14	-1.15	19.90	-1.21	22.76	-1.26	25.63

	MC in three explanatory variables															
	(10,10,10)		(11,11,11)		(12,12,12)		(13,13,13)		(14,14,14)		(15,15,15)		(16,16,16)		(17,17,17)	
α	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN
5	-0.95	12.56	-0.99	13.88	-1.03	15.07	-1.06	16.3	-1.09	17.57	-1.12	18.780	-1.15	20.08	-1.18	21.39
10	-1.10	17.47	-1.14	19.6	-1.18	21.32	-1.21	23.19	-1.24	24.91	-1.27	26.69	-1.3	28.48	-1.33	30.25
15	-1.19	21.85	-1.23	24	-1.26	26.19	-1.30	28.4	-1.33	30.64	-1.36	32.77	-1.39	35.03	-1.42	37.28
20	-1.25	25.44	-1.30	28.06	-1.33	30.46	-1.36	32.84	-1.40	35.4	-1.43	37.96	-1.45	40.56	-1.48	43.19
25	-1.30	28.41	-1.36	31.51	-1.40	34.2	-1.42	37.17	-1.45	40	-1.48	42.73	-1.51	45.67	-1.53	48.4

	MC in three explanatory variables															
	(18,18,18)		(19,19,19)		(20,20,20)		(21,21,21)		(22,22,22)		(23,23,23)		(24,24,24)		(25,25,25)	
α	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN	HLCIM	CN
5	-1.2	22.6	-1.22	23.86	-1.25	25.07	-1.27	26.47	-1.29	27.63	-1.31	28.77	-1.33	30.26	-1.34	31.31
10	-1.35	32.05	-1.4	33.9	-1.4	35.69	-1.42	37.34	-1.44	39.2	-1.46	40.82	-1.38	42.76	-1.49	44.25
15	-1.44	39.35	-1.47	41.86	-1.49	43.96	-1.51	45.87	-1.53	48.15	-1.55	50.32	-1.57	52.65	-1.58	54.51
20	-1.51	45.7	-1.53	48.15	-1.56	50.81	-1.57	53.42	-1.59	55.84	-1.61	58.33	-1.63	60.63	-1.65	63.28
25	-1.55	50.97	-1.58	54.15	-1.6	57.08	-1.62	59.76	-1.65	62.9	-1.66	65.63	-1.68	68.6	-1.7	71.54

High Leverage Collinearity- Influential Measure ($(\log\left(\frac{k_{(D)}}{k}\right)$), CN# Condition Number of X matrix, MC# Magnitude of Contamination

indicate the degree of multicollinearity at different sample sizes and different number of explanatory variables. In this simulation study, we considered different sample sizes that varied from 20, 100, 500 and 1000, and different number of explanatory variables, that is 3, 5 and 10. The maximum number of explanatory variables and sample sizes were chosen to be 10 and 1,000 because in real situation, we seldom encounter real data sets with more than these considered values. The same processes of simulation experiments as explained before for the three explanatory variables and sample size 100, were performed. Table 5 presents the LB and UB values of the HLCIM for different sample sizes and different number of explanatory variables. By increasing the number of collinear explanatory variables, the values of LB and UB in different sample sizes decreased. It is important to mention here that it is not an easy task to obtain the exact LB and UB values of the HLCIM. In this respect, we propose to use the extrapolation technique to obtain the values of LB and UB for different sample sizes between 20 and 1000 and explanatory variables between 2 and 10.

NUMERICAL EXAMPLE

A well-known data set, which is referred to for diagnosing influential observations is the Hawkins et al. (1984) data set. Hawkins et al. (1984) constructed an artificial three-predictor data set containing 75 observations with 14 influential observations, ten high leverage outliers (cases 1–10) and four high leverage points (cases 11–14). The D Group consists of these 14 high leverage points which is considered as suspected group of high leverage collinearity-enhancing observation. This data set contains 18.67 percent high leverage points (between 15 and 20 percent). Table 6 presents the values of the Diagnostic Robust Generalized Potential (DRGP-MVE) and MC for all three explanatory variables for the first 14 observations in this data set. The results of Table 6 signify these points as multiple high leverage points. To compute MC, first we needed to compute RD for a clean data set (the data set without these 14 multiple high leverage points which were normalized to have mean zero and standard deviation 1). The RD for each of the three explanatory variables X_1, X_2, and X_3 were equal to 3.73, 3.93 and 4.21, respectively.

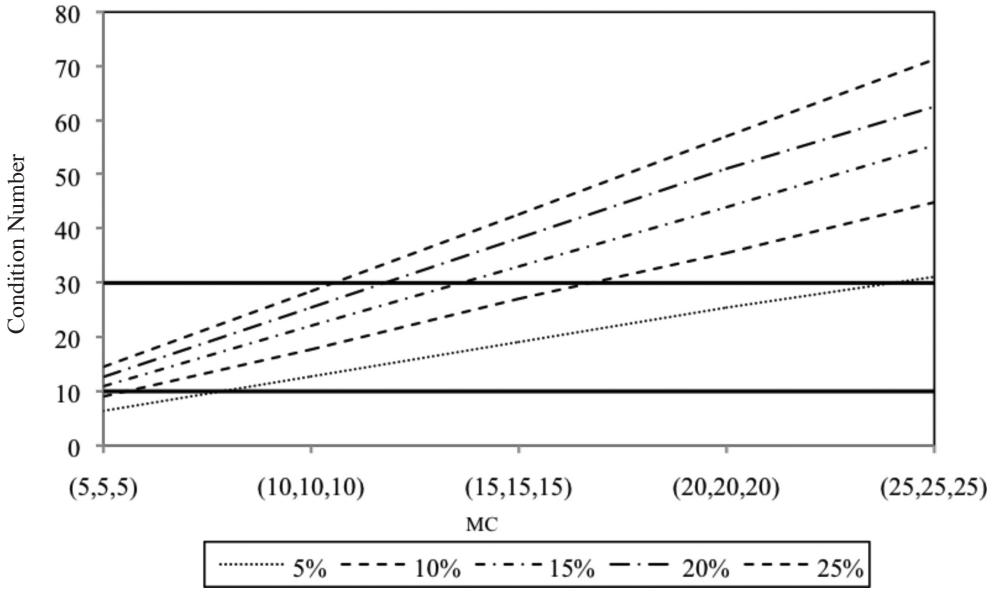

FIGURE 2. Condition Number (CN) against Magnitude of Contamination (MC) for three contaminated explanatory variables, n=100

TABLE. 5 LB and UB of HLCIM for different sample sizes and explanatory variables

explanatory variables no.	sample size							
	20		100		500		1000	
	LB	UB	LB	UB	LB	UB	LB	UB
3	-0.88	-1.30	-0.90	-1.34	-0.99	-1.44	-1.02	-1.58
5	-0.68	-1.17	-0.87	-1.30	-0.94	-1.42	-0.96	-1.49
10	-0.37	-0.85	-0.77	-1.25	-0.90	-1.37	-0.92	-1.41

LB# Lower Bound of HLCIM , UB# Upper Bound of HLCIM, HLCIM# High Leverage Collinearity Influential Measure ($\log\left(\frac{k_{(D)}}{k}\right)$),

TABLE 6. Diagnostic Robust Generalized Potential (DRGP) (MVE) and MC for Hawkins–Bradu–Kass data set

index	DRGP (MVE)(0.21)	MC for X_1	MC for X_2	MC for X_3
1	13.44	2.71	4.98	6.71
2	14.18	2.55	5.21	6.86
3	17.00	2.87	5.13	7.35
4	17.84	2.66	5.47	7.52
5	16.86	2.76	5.36	7.38
6	14.06	2.90	5.19	6.93
7	13.81	2.82	5.31	6.90
8	14.26	2.66	4.98	6.83
9	17.22	2.60	5.26	7.35
10	16.83	2.50	5.01	7.19
11	21.56	2.95	6.10	8.30
12	25.47	2.18	3.83	6.12
13	18.29	2.18	4.33	5.62
14	17.99	2.00	5.66	5.62

Thus, the MC was obtained by dividing the high leverage points in each explanatory variable to its corresponding RD. The average value of MC for X_1, X_2 and X_3 were equal to 2.81, 5.65 and 7.44, respectively. The average value of MC for the three explanatory variables was equal to 5.30, which was more than two for all three explanatory variables. Thus, it was another evidence to be sure that the contaminated points of this data were multiple high leverage points. To decide if multiple high leverage points may be the source of multicollinearity problems in this data set, the condition number of X matrix for the data set with and without high leverage points needed to be computed, and they were 12.42 and 1.18, respectively. Thus, it is obvious that the multicollinearity was due to the multiple high leverage points. To decide the degrees of multicollinearity, the simulation results in Table 5 had to be extrapolated. The LB and UB values of HLCIM when $n = 75$ and $p=3$ should be computed. These can be done easily by the extrapolation of the results for $n = 20$ and 100 in $p = 3$. Thus, the LB and the UB when $n = 75$ and $p = 3$ were -0.89 and -1.31 respectively. The computation of $\log\left(\frac{k_{(D)}}{k}\right)$ gave the value of HLCIM equaling to -1.02, where it lies between these two ranges which suggested a moderate multicollinearity. This result can be confirmed from the value of CN of the X matrix for the whole data set, which was equal to 12.42. The result of the simulation study confirmed that this data set had a moderate multicollinearty problem, which was due to the multiple high leverage points. Thus, the average contamination magnitude of 5.3 and 18.67 percent high leverage points may cause moderate multicollinearity in this data set.

CONCLUSION

Multicollinearity causes major interpretive problems in regression analysis, such as wrong sign problems, produces unstable and inconsistent estimates of parameters and insignificant regression coefficients, where in fact it is significant. Thus, it is very essential to investigate and detect the presence of multicollinearity to reduce its destructive effects on the regression estimates. It is now evident that high leverage points are a new prime source of multicollinearity. However, little work has been explored in this area. The main focus of this paper is to study the effect of different magnitude and different percentage of high leverage points on multicollinearity problems. In addition to that we investigated that in which extent; these high leverage points can cause multicollinearity. This paper also attempts to develop and introduce reliable cutoff points for High Leverage Collinearity Influential Measure (HLCIM) to diagnose different degrees of multicollinearity. Monte Carlo simulations were carried out to study these aims. The simulation indicated that there were four important factors that make high leverage points to be high leverage collinearity-enhancing observations. The factors are: the number of sample size, the number of contaminated explanatory variables, the magnitude of contamination and the percentage of high leverage points. It is interesting to note that the contaminated points, which exist in only one explanatory variable, cannot cause multicollinearity problems. High leverage collinearity-enhancing observations are those points in which their values are in large magnitude for at least two explanatory variables. The results also signify that when the high leverage points are in the same observations of two explanatory variables, by increasing the magnitude and the percentage of high leverage points, these

points tend to be high leverage collinearity-enhancing observations. Moreover, by increasing the percentage of high leverage points, the high leverage collinearity-enhancing observations can be detected at smaller MCs. The simulation experiments also show that the Lower Bound and Upper Bound of HLCIM corresponding to moderate and severe multicollinearity equals approximately to -0.90 and -1.34 in sample size equal to 100. By increasing the number of contaminated explanatory variables for fixed sample sizes, the value of LB and UB of HLCIM decreases. These values increase by increasing the number of sample sizes for fixed contaminated explanatory variables. Since the cutoff values for HLCIM that indicated moderate and severe multicollinearity were not easy to obtain, as an alternative, the extrapolation technique, which was acquired from the simulation results of Table 5 is recommended to compute LB and UB of HLCIM for a specific number of sample size and number of contaminated explanatory variables.

REFERENCES

Belsley, D.A. 1984. Demeaning conditioning diagnostics through centering (with comments). *The American Statistician* 38(2): 73-93.

Belsley, D.A. 1991. Conditioning Diagnostics - Collinearity and Weak Data in Regression. In *Probability and Mathematical Statistics*, New York: Wiley Series.

Belsley, D.A., Kuh, E. & Welsch, R.E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Habshah, M., Norazan, M.R. & Imon, A.H.M.R. 2009. The performance of Diagnostic-Robust Generalized Potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics* 36(5): 507-520.

Hadi, A.S. 1992. A new measure of overall potential influence in linear regression. *Computational Statistics & Data Analysis* 14: 1-27.

Hadi, A.S. 1988. Diagnosing collineariy-influential observations. *Computational Statistics & Data Analysis* 7: 143-159.

Hawkins, D.M., Bradu, D. & Kass, G.V. 1984. Location of several outliers in multiple regression data using elemental sets. *Technometrics* 26: 197-208.

Hoaglin, D.C. & Welsch, R.E. 1978. The Hat Matrix in regression and ANOVA. *Journal of American Statistical Association* 32: 17-22.

Hocking, R.R. & Pendelton, O.J. 1983. The regression dilemma. *Communications in Statistics-Theory and Methods* 12: 497-527.

Imon, A.H.M.R. 2002. Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies* 3: 207-218.

Kamruzzaman, M.D. & Imon, A.H.M.R. 2002. High leverage point: another source of multicollinearity. *Pakistan Journal of Statistics* 18: 435-448.

Maronna, R.A., Martin, R.D. & Yohai, V.J. 2006. *Robust Statistics Theory and Methods*. New York: Wiley & Sons.

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

Mason, C.H. & Perreault, jr. W.D. 1991. Collinearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research* XXVIII (August): 268-280.

Moller, S.F., Frese, J.V. & Bro. R. 2005. Robust Methods for multivariate data analysis. *Journal of Chemometrics* 19: 549-563.

Montgomery, D.C., Peck, E.A. & Viving, G.G. 2001. *Introduction to linear regression Analysis*. 3th ed. New York: John Wiley and Sons.

Neter, J., Kutner, M.H., Wasserman W. & Nachtsheim, C.J. 2004. *Applied Linear Regression Models*. New York: MacGRAW-Hill/Irwin.

Rosen, D.H. 1999. The Diagnosis of Collinearity: A Monte Carlo Simulation Study, Department of Epidemiology. PhD thesis School of Emory University.

Rousseeuw, P.J. 1985. Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, edited by Reidel Dordrecht *B:* 283-297.

Schindler, J.S. 1986. Regression Diagnostics: Mechanical and Structural Aspects of Collinearity. PhD thesis Department of Biostatistics. University of North Carolina at Chapel Hill.

Sengupta, D. & Bhimasankaram, P. 1997. On the roles of observations in collineariy in the linear model. *Journal of American Statistical Association* 92(439): 1024-1032.

Stinnett, S.S. 1993. Collinearity in Mixed Models. PhD thesis Department of Biostatistics, University of North Carolina at Chapel Hill.

Habshah Midi
Faculty of Science
Universiti Putra Malaysia
43400 Serdang, Selangor D.E.
Malaysia

Habshah Midi, Arezoo Bagheri*
Institute For Mathematical Research (INSPEM)
Universiti Putra Malaysia
43400 Serdang, Selangor D.E.
Malaysia

A.H.M. Rahmatullah Imon
Department of Mathematical Sciences
Ball State University, Muncie
IN 47306, USA

*Corresponding author; email: abagheri_000@yahoo.com

Received: 5 February 2010
Accepted: 11 November 2010