Journal of Applied Sciences

ISSN 1812-5654

science alert

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

Journal of Applied Sciences 10 (18): 2086-2093, 2010
ISSN 1812-5654

The Effect of Collinearity-influential Observations on Collinear Data Set: A Monte Carlo Simulation Study

[1]A. Bagheri, [1]Habshah Midi 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, the effect of different patterns of high leverages on the classical multicollinearity diagnostics and collinearity-influential measure is investigated. Specifically the investigation is focus on in which situations do these points become collinearity-enhancing or collinearity-reducing observations. Both the empirical and the Monte Carlo simulation results, in collinear data sets indicate that when high leverages exist in just one explanatory variable or when the values of the high leverages are in different positions of the two explanatory variables, these points will be collinearity-reducing observations. On the other hand, these high leverages are collinearity-enhancing observations when their values and positions are the same for the two collinear explanatory variables.

Key words: High leverage points, multicollinearity, diagnostic methods, condition number, collinearity-influential measure

INTRODUCTION

Nonorthogonality of explanatory variables or near-linear dependency between two or more explanatory variables is called Multicollinearity. The presence of Multicollinearity has some destructive effects on regression analysis such as prediction inferences and estimations. Consequently, the validity of parameter estimations becomes questionable (Montgomery *et al*., 2001; Kutner *et al*., 2004; Chatterjee and Hadi, 2006; Midi *et al*., 2010). Kamruzzaman and Imon (2002) and Montgomery *et al*. (2001) pointed out that there are different sources of multicollinearity such as 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 there is no statistical test for the presence of this problem in the data set. Therefore, a diagnostic method can be used to indicate the existence and extent of multicollinearity in a data set. Belsley *et al*. (1980) proposed Condition Number (CN) of X matrix as a very practical multicollinearity diagnostic method which may be obtained from the singular-value decomposition of the (n×p) 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 indication of the degree of multicollinearity in the data set. CN for X matrix between 10 and 30 has been recommended by him as an indicator of 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. Many studies have been devoted to this issue (Mason and Perreault Jr., 1991; Rosen, 1999).

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 and Pendelton, 1983; Moller *et al*., 2005) can affect classical multicollineaity diagnostics methods. These points may be a new source of multicollinearity by the usage of classical multicollinearity diagnostics methods which have been introduced by Kamruzzaman and Imon (2002). According to Hadi (1988), the high leverage points or outliers in the X-direction may be collinearity-influential observations. He noted that the collinearity-influential observations are usually points with high leverages while all high leverage points are not collinearity-influential observations. It is worth mentioning that high leverage points can be collinearity-

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

influential observations according to the classical multicollinearity diagnostics methods. Hadi (1988) defined a collinearity-influential measure based on condition number of X matrix. This measure not only suffers from defining a practical cutoff point but also the lack of symmetry which is due to the additive change in condition number of X matrix. Sengupta and Bhimasankaram (1997) pointed out the weakness of Hadi's measure and proposed a new practical collineaity influential measure.

Yet, little attention has been devoted to the role of individual cases in collinearity of explanatory variables especially in the collinear data sets (Midi *et al*., 2010). Furthermore, there is a lack of investigation in the literature on high leverage points that cause multicollinearity problems.

It is necessary to study the effect of the high leverage collinearity-influential observations on the most applicable multicollinearity diagnostics such as the Collinearity-Influential Measure (CIM) and CN (Midi *et al*., 2010). In this way, we can investigate the change in degree of multicollinearity caused by the high leverage points in collinear data set. Unfortunately, there is no direct technique to investigate the effect of high leverage pints on collinearity pattern of a collinear data set. Insight is gained only by simulation experiences and by real data sets (Rosen, 1999; Midi *et al*., 2010).

Before proceeding to the simulation study, diagnostic methods of high leverage points will be reviewed briefly. Moreover, the effect of high leverage points in colinearity pattern of a real well-known collinear data set will be investigated. In addition, the Monte Carlo simulation study will be carried out to confirm the result of real data.

MATERIALS AND METHODS

Collinearity-influential measures: Let define a regression model as:

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

where, Y is an (n×1) vector of response or dependent variables, X is an (n×p) matrix of predictors, n×p, β is a (p×1) vector of unknown finite parameters to be estimated and ε is an (n×1) vector of random errors. We let the jth column of the X matrix be denoted as X_j, therefore, $X = [X_1, X_2,....,X_p]$. Additionally, we defined multicollinearity in terms of the linear dependence of the columns of X, i.e., whereby the vectors of $X_1, X_2, \ldots, X_p$ are linearly dependent if there is a set of constants $t_1, t_2, \ldots, t_p$, that are not all zero, such as:

$$\sum_{j=1}^{p} t_j X_j = 0 \quad (2)$$

We face severe multicollinearity problem, if Eq. 2 holds exactly. If Eq. 2 holds approximately, the problem of moderate multicollinearity is said to exist.

A very practical multicollinearity diagnostic method can be obtained from the singular-value decomposition of the (n×p) X matrix which has been proposed by Belsley *et al*. (1980). The X matrix can be decomposed as:

$$X = UDV^T \quad (3)$$

where, U is the (n×p) matrix in which the columns that are associated with the p non-zero eigen values of X^TX are the eigenvectors, V is the (p×p) matrix of eigen vectors of, X^TX, $U^TU = I$, $V^TV = I$ and D is a (p×p) diagonal matrix with non-negative diagonal elements k_j, j = 1,2,…,p which is called singular-values of X. Furthermore, they defined the condition indices of the X matrix as:

$$k_j = \frac{\lambda_{max}}{\lambda_j}; \quad j = 1,2,...,p \quad (4)$$

where, $\lambda_1, \lambda_2,....,\lambda_p$ are the singular values of X matrix. It is noticeable that the largest value of k_j can be defined as Condition Number (CN) of X matrix. The explanatory variables should be scaled to have the same length before calculating the condition indices to make them comparable from one data set to another. Moreover, scaling the independent variables prevents the eigen analysis of X matrix to be dependent on the variables units of measurements.

Furthermore, he defined a measure for the influence of the ith row of X matrix on the condition number as:

$$\delta_i = \frac{k_{(i)} - k}{k}; \quad i = 1,....,n \quad (5)$$

where, $k_{(i)}$ can be computed from the eigen values of $X_{(i)}$ when the ith row of X matrix has been deleted. Hadi specified 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, Hadi's measure is not practical because he did not provide any cutoff points. The decision as to how large the value of δ_i should be, is solely depend on the researcher's judgment on the magnitude of this value.

Sengupta and Bhimasankaram (1997) pointed out the weakness of Hadi's measure is in the lack of symmetry, which is due to the additive change in k. To overcome this problem, they proposed:

$$l_i = \log\left(\frac{k_{(i)}}{k}\right); \qquad i = 1,.....,n \tag{6}$$

As a Collinearity-Influential Measure (CIM) for each row of observations. Although Sengupta and Bhimasankaram (1997) didn't propose any specific cutoff point for CIM, they introduced some easily computable lower bound and upper bound values for this new collinearity- influential measure. It is important to note that high leverage points can hide and induce multicollinearity pattern in two different situations: when $\frac{k_{(i)}}{k} > 1$ and $k_{(i)} > k$, then $\log\left(\frac{k_{(i)}}{k}\right) > 0$ and when $0 < \frac{k_{(i)}}{k} < 1$ and $k_{(i)} < k$, then $\log\left(\frac{k_{(i)}}{k}\right) < 0$. In the first situation, the degree of multicollinearity increases due to the characteristics of high leverages which hide the multicollinearity pattern. Thus, the high leverage points are referred as collinearity-reducing observation. Otherwise, the deletion of high leverage points may reduce the degree of multicollinearity. Hence, in the second situation the high leverage points are referred as collinearity-enhancing observation.

High leverage diagnostics methods: There are different types of outlier diagnostics methods for univriate and multivariate regression models (Belsley, 1991; Kutner *et al*., 2004; Wilcox, 2005). One of the practical outlyingness diagnostics 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:

$$W = X\left(X^TX\right)^{-1}X^T$$

The most widely used cutoff point of the hat matrix is the twice-the-mean-rule (2k/n) by Hoaglin and 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:

$$p_{ii} = \frac{w_{ii}}{1 - w_{ii}} \tag{7}$$

where, $w_{ii} = x_i^T(X^TX)^{-1}x_i$ is the diagonal element of W and the ith, diagonal potential p_{ii} can be defined as $p_{ii} = x_i^T(X_{(i)}^TX_{(i)})^{-1}x_i$ where $X_{(i)}$ is the data matrix X without the ith row. He proposed a cutoff point for potential values p_{ii} as Median(p_{ii}) +c MAD(p_{ii}) (MAD-cutoff point) where is normalized Median Absolute Deviation defined by:

$$MAD(\theta) = \text{median}\left|p_{ii} - \text{median}(p_{ii})\right|/0.6745 \qquad i = 1,2,...,n$$

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:

$$p_{ii}^* = \begin{cases} w_{ii}^{(-D)} & \text{for } i \in D \\ \dfrac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} & \text{for } i \in R \end{cases} \tag{8}$$

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. But, since 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) (Bagheri *et al*., 2009). We refer this method as the DRGP(MVE). The set D (deletion set) in generalized potentials method in Eq. 8 is defined basedon the points which RMD-MVE exceeds Median (RMD-MVE)+3MAD(RMD-MVE). Rousseeuw (1985) introduced RMD-MVE as:

$$RMD_i = \sqrt{\left(X - T_R(X)\right)^T C_R(X)^{-1}\left(X - T_R(X)\right)} \quad \text{for} \quad i = 1,.....,n \tag{9}$$

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

Body fat data set: Here, the effect of high leverage points on a collinear data set which was introduced by Kutner *et al*. (2004) has been investigated. Body fat data set is a three explanatory variables data set with 20 observations. Triceps skinfold thickness (X_1), thigh circumference (X_2) and midarm circumference (X_3) are its three explanatory variables. Kutner *et al*. (2004) mentioned that this data set has multicolinearity problem. Table 1 presents the high leverage points for the body fat data.

Table 1: High leverage diagnostics methods for body fat data set

Index	w_{ii} (0.4000)	DRGP(MVE) (0.3678)	Index	w_{ii} (0.4000)	DRGP (MVE) (0.3678)
1	0.3412	0.4360	11	0.1394	0.1524
2	0.1565	0.0806	12	0.1093	0.1314
3	0.4404	0.7875	13	0.2136	0.3892
4	0.1124	0.1379	14	0.1881	0.3082
5	0.3611	0.3668	15	0.3483	0.5414
6	0.1315	0.1706	16	0.1144	0.1131
7	0.1943	0.2074	17	0.1253	0.1247
8	0.1642	0.1187	18	0.2283	0.2759
9	0.1928	0.1889	19	0.1324	0.1116
10	0.2405	0.1428	20	0.0660	0.0631

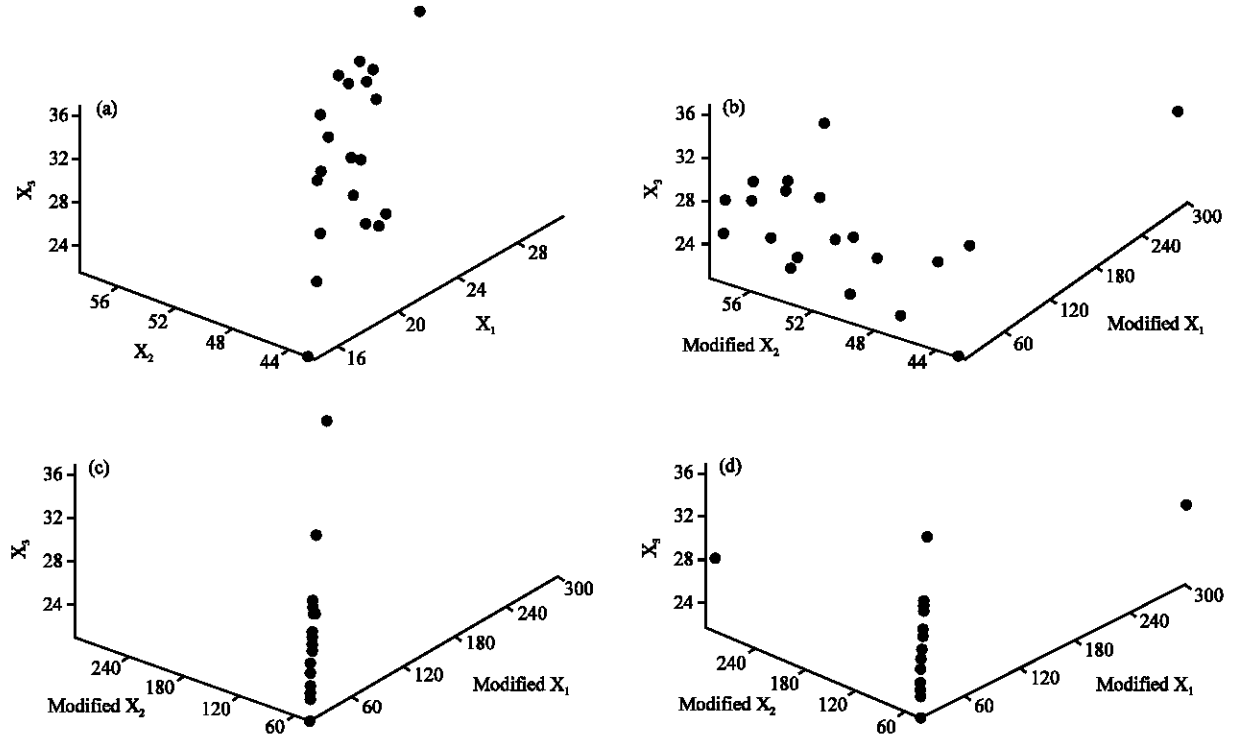

Fig. 1: Scatter plot of original and modified body fat data set, (a) original data set, (b) modified X_1, (c) modified X_1 and X_2 in the same positions and (d) modified X_1 and X_2 in different positions

Here, the explanatory variables have been scaled to protect the condition number from being dominated by some large explanatory variables. Scaling will prevent the eigen analysis to be dependent on the variables units of measurements.

To compute the condition number of X matrix, the explanatory variables have been scaled following the scaling method by Stewart (1987), in the following form:

$$x'_{ij} = \frac{x_{ij}}{\|X_j\|} \quad i = 1,...,p, \quad j = 1,...,n \tag{10}$$

For alternative scaling methods one can refer to Stewart (1987) and Hadi (1988).

To study the effect of high leverage points on different collinearity patterns, the original collinear data is modified to have high leverage point just in one explanatory variable, the same values and positions of high leverage points in the two explanatory variables and the same value of high leverage points but in different positions of the two explanatory variables. The first collinearity pattern is created by replacing the first observation of the first explanatory variables with 300. In the second situation, the first observations of X_1 and X_2 are substituted with the same value equals to 300. Finally, in the third situation, the first observation of X_1 and the last observation of X_2 are replaced with 300. Figure 1 shows the matrix plot of the original and modified Body fat data set.

Table 2 and 3 exhibit the effect of high leverage points on the OLS estimates in the collinear Body fat data set.

Table 2: Parameter estimation for original and modified in X_1 data set

	Original data set				Modified X_1			
Coefficient	Value	SE	t-value	p-value	Value	SE	t-value	p-value
Intercept	117.0847	99.7824	1.1734	0.2578	-24.9715	7.2794	-3.4304	0.0034
b1	4.3341	3.0155	1.4373	0.1699	-0.0068	0.0103	-0.6660	0.5149
b2	-2.8568	2.5820	-1.1064	0.2849	0.8261	0.1202	6.8697	0.0000
b3	-2.1861	1.5955	-1.3701	0.1896	0.1146	0.1664	0.6885	0.5010
F-value	21.5200				19.1000			
p-value	0.0000				0.0000			

Table 3: Parameter estimation for modified in x_1 and x_2 in the same positions and modified in x_1 and x_2 in the different positions

	Modified X_1 and X_2 (same position)				Modified X_1 and X_2 (different position)			
Coefficient	Value	SE	t-value	p-value	Value	SE	t-value	p-value
Intercept	-134.1269	26.2826	-5.1033	0.0001	14.1681	8.8500	1.6009	0.1289
b1	-3.3597	0.5795	-5.7975	0.0000	-0.0272	0.0190	-1.4298	0.1720
b2	3.6645	0.6374	5.7490	0.0000	0.0186	0.0211	0.8819	0.3909
b3	1.8780	0.3374	5.5664	0.0000	0.2143	0.3253	0.6588	0.5194
F-value	13.6400				1.1550			
p-value	0.0001				0.3574			

Table 4: The effect of different modification on collinearity-influential observation and condition No. for MC = 100 in body fat data set

Status of high leverage points	CIM	CN
Original data set	-	23.6208
Modified data set (HL in X_1)	0.5890	14.7708
Modified data set (HL in the same position of X_1 and X_2)	-1.4331	33.2444
Modified data set (HL in the different position of X_1 and X_2)	1.1927	3.1717

HL: High leverages

Table 4 presents the effect of adding high leverage points on CN and CIM.

Monte Carlo simulation study: Here, we report a Monte Carlo simulation study that is designed to investigate the effect of different sample sizes and different magnitude and percentage of high leverage points on CIM in collinear data sets. Following the idea of Lawrence and Arthur (1990), three explanatory variables were generated as follows:

$$X_{ij} = (1-\rho^2)Z_{ij} + \rho Z_{i4} \quad i = i,....,n; \quad j = 1,.....,3 \qquad (11)$$

where, the X, Z are independent standard normal random numbers. The value of ρ^2 represents the correlation between the two explanatory variables.

In the simulated data sets, the value of ρ^2 was chosen to be equals to 0.95 which results in high collinearity between explanatory variables that created collinear data sets. Different percentage of high leverage points has been added to the explanatory variables. The level of high leverage points (α) is varied from zero to 25 % and four samples of size 20, 60, 100 and 300 were considered. The magnitude of high leverage points has been varied from 20, 50, 100 and 300. Three different situations of adding high leverage points to the data set were carried out in the manner described earlier similar to the Body fat data example.

Table 5: CN and CIM for different percentage and magnitude of high leverage points and n = 20

	Contamination					
	Pattern 1		Pattern 2		Pattern 3	
	CIM	CN	CIM	CN	CIM	CN
MC = 20						
5	0.3232	32.4727	-2.1372	381.2066	2.9053	2.4575
10	0.3278	32.6095	-2.4889	546.7006	2.9226	2.4620
15	0.3350	32.6188	-2.7026	682.5362	2.9444	2.4586
20	0.3395	32.7045	-2.8605	806.0937	2.9409	2.5463
25	0.3502	32.8273	-2.9776	919.3104	2.9861	2.5459
MC = 50						
5	0.3219	32.4232	-3.0544	952.2812	2.9321	2.3872
10	0.3276	32.4695	-3.4089	1365.9475	2.9492	2.3830
15	0.3356	32.5210	-3.6234	1708.8823	2.9714	2.3904
20	0.3398	32.6363	-3.7771	2010.7567	2.9642	2.4870
25	0.3472	32.7357	-3.9042	2309.3622	3.0069	2.4928
MC = 100						
5	0.3236	32.3676	-3.7487	1906.4222	2.9384	2.3717
10	0.3308	32.4406	-4.1043	2747.1931	2.9566	2.3735
15	0.3369	32.5423	-4.3179	3433.1877	2.9839	2.3708
20	0.3390	32.7063	-4.4730	4041.3609	2.9709	2.4704
25	0.3487	32.7562	-4.5932	4608.3793	3.0148	2.4699
MC = 300						
5	0.3261	32.2546	-4.8484	5720.9228	2.9412	2.3617
10	0.3285	32.4632	-5.2068	8261.9747	2.9592	2.3612
15	0.3328	32.6117	-5.4179	10284.9360	2.9817	2.3652
20	0.3429	32.6865	-5.5722	12173.5350	2.9746	2.4676
25	0.3493	32.6810	-5.6934	13822.2540	3.0169	2.4628

#: Percentage of high leverage points; MC: Magnitude of high leverage points; CIM: Collinearity-influential measure

The first and the second contamination patterns were created by replacing the first 100% of the observations of X_1 and the first 100% of the observations of X_1 and X_2 with certain magnitude of high leverage points, respectively. The last pattern was generated by substituting certain magnitude of high leverage points to

Table 6: CN and CIM for different percentage and magnitude of high leverage points and n = 300

	Contamination					
	Pattern 1		Pattern 2		Pattern 3	
	CIM	CN	CIM	CN	CIM	CN
MC = 50						
5	0.2284	30.0635	-2.2217	348.4678	2.7951	2.3106
10	0.2260	30.1688	-2.5792	498.7601	2.7903	2.3277
15	0.2253	30.2194	-2.7962	620.2199	2.7783	2.3617
20	0.2227	30.3520	-2.9558	728.9733	2.7633	2.4093
25	0.2196	30.4736	-3.0842	829.5418	2.7500	2.4531
MC = 50						
5	0.2312	29.9942	-3.1366	870.4778	2.8328	2.2263
10	0.2295	30.0662	-3.4974	1249.5077	2.8165	2.2669
15	0.2273	30.1820	-3.7151	1555.9657	2.8028	2.3072
20	0.2252	30.2877	-3.8755	1829.3269	2.7842	2.3585
25	0.2227	30.3891	-4.0042	2082.1554	2.7702	2.4043
MC = 100						
5	0.2330	29.9293	-3.8298	1740.2711	2.8416	2.2057
10	0.2302	30.0586	-4.1916	2502.8817	2.8240	2.2513
15	0.2286	30.1460	-4.4096	3116.1365	2.8082	2.2938
20	0.2267	30.2408	-4.5706	3665.5639	2.7898	2.3440
25	0.2239	30.3539	-4.7003	4176.8323	2.7752	2.3920
MC = 300						
5	0.2326	29.9434	-4.9287	5222.4572	2.8458	2.1961
10	0.2310	30.0314	-5.2905	7509.9843	2.8289	2.2401
15	0.2286	30.1053	-5.5101	9354.2623	2.8117	2.2831
20	0.2268	30.2084	-5.6698	10992.2543	2.7947	2.3330
25	0.2241	30.3516	-5.7990	12534.2039	2.7791	2.3826

#: Percentage of high leverage points; MC: Magnitude of high leverage points; CIM: Collinearity-influential measure

the first $100(\alpha/2)$ percent observations of X_1 and to the last 100 $(\alpha/2)$ percent observations of X_2 to have high leverage points in different positions of these two explanatory variables. Moreover, to compute the values of CN for X matrix, the generated explanatory variables have been scaled according to equation of Eq. 11. In each simulation run, there were 10,000 replications.

The values of CN and CIM for the three contamination patterns, different magnitude of high leverage points and for small sample (n = 20) and large sample (n = 300) are displayed in Table 5 and 6, respectively.

DISCUSSION

The results that we have obtained in the previous section will be discussed in this section. At first we discuss the results of the numerical examples. The result of Table 1 pointed out that the hat matrix can detect case 3 while DRGP (MVE) indicates three more cases 1, 3 and 15 as high leverage points. To see the effect of these high leverage points on classical diagnostics methods, the CN of X matrix for with and without high leverage points have been calculated.

The CN of X matrix in the presence of high leverage points in the data set is equals to 23.6208. After these high leverage points (1, 3 and 15) are removed from the data set, the CN value is equals to 32.3298 which indicates that this data set has severe multicollinearity problem. This result also pointed out that the leverage points are not the cause of multicollinearity. They only reduce the degree of multicollinearity from severe to moderate.

According to the Fig. 1a, there is an obvious linear relationship between all three explanatory variables (Montgomery *et al*., 2001; Kutner *et al*., 2004). By adding high leverage points in just X_1, the multicollinearity pattern of the X_1 and X_2 has been ruined (Fig. 1b) while by adding high leverage point to both X_1 and X_2 again multicollinearity appeared between these two explanatory variables (Fig. 1c). The plot of Fig. 1d seems to suggest that the multicollinearity pattern of the data is being masked.

If the high leverage points are added in different positions of the explanatory variables, Table 2 and 3 exhibit the effect of high leverage points on the OLS estimates in this collinear data set. It can be observed from Table 2 that when multicollinearity exists in this data set, the p-value of the F-test reveals that there is a linear relationship between the explanatory variables while none of the t-values of the explanatory variables are significant. This is another indicator of the presence of multicollinearity in this data set (Kutner *et al*., 2004; Chatterjee and Hadi, 2006). When modification is done only on X_1, the F-test is still significant but again t-values of X_1 and X_3 are not significant. Several interesting points can be seen from Table 3. When modifying both X_1 and X_2 with the same values and positions, the F-value and all the t-values are significant. These results incorrectly indicate that there is no multicollinearity problem in the data where in fact multicollinearity exist. The reason for this misleading result is that the collinearity of the explanatory variables in this data set has been masked by the presence of the high leverage points in the same values and positions of both X_1 and X_2. In this study, the effect of different pattern of high leverage points on the classical multicollinearity diagnostics and collinearity-influential measure is investigated. On the other hand, in the situation when modification on X_1 and X_2 are with different positions, the F-test and all of the t-tests are not significant that indicates that there is no linear relationship between the dependent and independent variables in regression model of this data set. Thus the presence of high leverage points on the collinear data set, can sometimes hide or increase the significance of the t-test for parameter estimation of a regression model.

It can be observed from Table 4 that the value of CN of X matrix reveals the presence of moderate multicollinearity problem in the original data set.. It is

worth mentioning that this data set suffers from severe multicollinearity (for comprehensive discussion method, refer to Kutner *et al*. (2004) according to the value of the Variance Inflation Factor (VIF). Since scaling is applied to the explanatory variables in the computation of CN of X matrix for this data set, it removes some of the ill-conditioning problem of the data set (Hadi, 1988; Montgomery *et al*., 2001; Midi *et al*., 2010). For the modification only on X_1 or in different positions of X_1 and X_2, these high leverage points are collinearity reducing observations evident by the CIM values which become positive and the smaller values of CN compared with the CN value of the original data set. It is interesting to note that modifying X_1 and X_2 in the same position, cause these points to be collinearity enhancing observations evident by the negative value of CIM and the larger values of CN compared with the original data. It is worth mentioning that the value of CIM has been computed from in Eq. 7 where i is the observation with high leverage which influence the multicollinearity pattern of the data.

Finally, we discuss the results obtained from the simulations which carried out to investigate whether these results confirm the conclusion of the real data set. Due to space limitations, we do not show all the simulation results and the results for small sample (n = 20) and large sample (n = 300) have been presented. The other results are consistent. The values of CN for X matrix without high leverage points for sample sizes 20 and 300 are equal to 44.3979 and 37.7396, respectively which indicate the existence of severe multicollinearity problem in the simulated data sets. Let us focus our attention to Table 5 and 6 for the contamination pattern 1. The values of CN for contaminated data are smaller than the uncontaminated data. However, it still indicates severe multicollinearity problem. It is worth mentioning that for sample of size 300, at 5% high leverage points and magnitude of contaminations 50, 100 and 300, the values of CN are very close to 30. The positive values of CIM also confirm that when the high leverage points are in only one explanatory variable, they are collinearity-reducing observations for collinear data sets. Nonetheless, for the contamination pattern 3, the scenarios change dramatically. The added high leverage points in different positions of X_1 and X_2 cause the collinear simulated data sets to be non-collinear as exhibited by small values of CN and positive values of CIM. Thus, the added high leverage points according to contamination pattern 3 are collinearity-reducing observations. On the other hand, when high leverage points are added to the collinear data sets according to the contamination pattern 2, the values of the CN become much larger than the CN of the uncontaminated data. It is evident from Table 5 and 6 that these points become collinearity-enhancing as its CN and CIM values become large and negative large, respectively. Therefore, when the same values of high leverage points are added to the same positions of two explanatory variables for collinear data sets, they will increase the multicollinearity problem of the data sets.

CONCLUSIONS

The main focus of this study is to investigate the effect of different magnitude and different percentage of high leverage points on the Collinearity-Influential Measure (CIM) of a collinear data. This paper also attempts to investigate the effect of high leverage points on the OLS estimates of collinear explanatory variables. The numerical example and Monte Carlo simulation study reveal that high leverage points can induce or reduce the multicollinearity pattern of a collinear data set. Furthermore, sometimes the high leverage points and multicolinearity mask each other's effects which lead researchers to rely on misleading results. To be high leverage collinearity-influential observations in collinear data set, the magnitude and percentage of high leverage points are essential factors, regardless of the sample size. The high leverage points, which exist in only one collinear explanatory variable, reduce the collinearity between these explanatory variables. The results also signify that when the same values of high leverage points are in the same positions for two collinear explanatory variables, by increasing the magnitude and the percentage of high leverage points, they increase the degree of collinearity among the explanatory variables. Moreover, when the same values of high leverage points are in different positions of the two collinear explanatory variables, these points reduced the collinearity between these explanatory variables.

REFERENCES

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

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

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, ISBN-10 0471746966.

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 Statist., 36: 507-520.

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

Hadi, A.S., 1992. A new measure of overall potential influence in linear regression. Comput. Stat. Data Anal., 14: 1-27.

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

Hocking, R.R. and O.J. Pendelton, 1983. The regression dilemma. Comm. Stat. Theory Meth., 12: 497-527.

Imon, A.H.M.R., 2002. Identifying multiple high leverage points in linear regression. J. Statist. Stud., 1: 207-218.

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.

Lawrence, K.D. and J.L. Arthur, 1990. Robust Regression: Analysis and Applications. Marcel Dekker Inc., New York, ISBN: 0-8247-8129-5, pp: 287.

Mason, C.H. and W.D. Perreault Jr., 1991. Collinearity, power and interpretation of multiple regression analysis. J. Market. Res., 28: 268-280.

Midi, H., A. Bagheri and A.H.M.R. Imon, 2010. The application of robust multicollinearity diagnostic method based on robust coefficient determination to a non-collinear data. J. Applied Sci., 10: 611-619.

Moller, S.F., J.V. Frese and R. Bro, 2005. Robust methods for multivariate data analysis. J. Chemometr., 19: 549-563.

Montgomery, D.C., E.A. Peck and G.G. Vining, 2001. Introduction to Linear Regression Analysis. 3rd Edn., Jon Wiley and Sons, New York, ISBN-10: 0471315656.

Rosen, D.H., 1999. The diagnosis of collinearity: A monte carlo simulation study. Ph.D. Thesis, Department of Epidemiology, School of Emory University.

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

Sengupta, D. and P. Bhimasankaram, 1997. On the roles of observations in collineariy in the linear model. J. Am. Stat. Assoc., 92: 1024-1032.

Stewart, G.W., 1987. Collinearity and least squares regression. Statist. Sci., 2: 68-84.

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