Professor Habshah MIDI, PhD
Department of Mathematics, Faculty of Science / Laboratory of Computational Statistics and Operations Research, Institute for Mathematical Research University Putra, Malaysia
Assistant Professor Arezoo BAGHERI, PhD
National Population Studies & Comprehensive Management Institute, Tehran, Iran
E- mail: abagheri@psri.ac.ir

ROBUST MULTICOLLINEARITY DIAGNOSTIC MEASURES BASED ON MINIMUM COVARIANCE DETERMINATION APPROACH

Abstract. *The classical multicollinearity diagnostic measures are not resistant to high leverage points since their formulation are based on eigen analysis of classical correlation matrix that is very sensitive to the presence of these leverages. The existing robust multicollinearity diagnostics also are not able to diagnose the variables which are collinear to each other. In this paper, we proposed robust multicollinearity diagnostic measures based on the Minimum Covariance Determination (MCD), which is a highly robust estimator of multivariate location and scatter. The results of numerical example and simulation study confirmed the merit of our new proposed robust multicollinearity diagnostic measures.*

Keywords: *Multicollinearity, Condition number, Variance Inflation Factor, Variance Decomposition Properties, high leverage points, Minimum Covariance Determination Method.*

JEL Classification: C15, C39, C63

1. INTRODUCTION

High leverage points that fall far from the majority of the explanatory variables are another source of multicollinearity, which is a linear relationship between the explanatory variables (Kamruzzaman & Imon 2002). These points are considered as good or bad leverage points based on whether they follow the same regression line as the other data in the data set or not. Furthermore, collinearity-influential observations are the observations that 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). Additionally, high leverage points that exist in different positions of several explanatory variables may be

collinearity-reducing observations and decrease the degree of multicollinearity for explanatory variables in collinear data sets (Bagheri & Habshah 2011; Bagheri et al. 2010). Moreover, in non-collinear data sets, high leverage points in the same positions of different explanatory variables cause multicollinearity problems and become collinearity- enhancing observations (Bagheri et al. 2012; Midi et al. 2010). At the same time, multicollinearity does have some destructive effects on the statistical inferences of multiple regression, for instance, causing least squares regression estimates to be unstable, inconsistent and have large variances. Another important effect of multicollinearity for the data set is when correlation matrix (X) is singular or near-singular. Multicollinearity can be detected by applying different classical diagnostic measures such as the condition number (CN), Variance Inflation Factor (VIF) and Variance Decomposition Properties (VDP) which can be computed based on the eigen value and eigen vectors of the correlation matrix (X) (Montgomery 2001). Nonetheless, all of these classical multicollinearity diagnostic measures are sensitive to the presence of high leverage points. Hence, it is very crucial to use robust diagnostic measures that are resistant to multicollinearity to avoid misleading conclusion in regression analysis. Furthermore, the existing robust multicollinearity diagnostic methods may not be able to diagnose the collinear variables (Bagheri and Habshah 2011,Midi et al. 2010). Rousseeuw (1985) introduced the Minimum Covariance Determinant (MCD) method, which aims to find h observations out of n observations which covariance matrix has the lowest determinant. It is important to note that the MCD estimators of location and scatter have high breakdown value while its efficiency is generally low. Several researches noted that one-step reweighting could increase the efficiency of MCD (Croux and Haesbroeck 1998; Lopuha¨a 1999, Rousseeuw and Van Driessen 1999). For many years, the Minimum Volume Ellipsoid (MVE) proposed by Rousseeuw (1985) was preferred over the MCD, because of its slightly better computational efficiency when using a resampling algorithm. However, many users prefer to use MCD as a robust estimator of location and scatter after the establishment of the fast-MCD algorithm by Rousseeuw and Van Driessen 1999. The basic ideas of Fast-MCD are in the inequality involving order statistics, determinants, and techniques, which we call selective iteration and nested extensions. For small datasets, FAST-MCD typically finds the exact MCD, whereas for larger datasets it gives more accurate results than the existing algorithms and is faster by orders of magnitude.

Since the classical diagnostic measures are not resistant to high leverage points, we take the initiative to develop robust multicollinearity diagnostic measures to remedy this problem. In this paper, we develop several robust multicollinearity diagnostic measures such as the Robust Condition Number (RCN), the Robust Variance Inflation Factor (RVIF) and the Robust Variance Decomposition Properties (RVDP). A one-step reweighted MCD is incorporated into the establishment of these robust multicollinearity diagnostic measures. The paper proceeds as follows. Section 2

reviews the classical multicollinearity diagnostic measures. The MCD is presented in Section 3. The application of MCD on robust multicollinearity diagnostic measures is presented in section 4. Finally, section 5 focuses on a well known collinear data set and the Monte Carlo simulation study. Some concluding remarks are offered in section 6.

2. CLASSICAL MULTICOLLINEARITY DIAGNOSTIC MEASURES

Consider the following regression model:

$$y = x\beta + \varepsilon \quad (1)$$

where y is an $n\times1$ vector of response or dependent variables, X is an $n\times p$ $(p = k + 1)$ design matrix (the number of independent variables predicting y is equal to k) , β is a $p \times1$ vector of unknown parameters and ε is an $n\times1$ vector of random errors with distribution $N[0, \sigma^2 I_n]$.

There is no statistical test for detecting the presence of multicollinearity in the data set, since it is a data problem. In this situation, a diagnostic measure can replace a statistical test to indicate the existence and extent of multicollinearity in a data set. 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 $X_1, X_2 ,\ldots,X_p$ are linearly dependent if there is a set of constants $t_1, t_2,\ldots,t_p$, that are not all zero, such as:

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

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

By applying unit length scaling introduced by (Montgomery 2001) or the correlation transformation indicated by Kutner et al. 2004 as:

$$Z_{ij} = \frac{X_{ij}-\bar{X}}{S_j} \qquad i=1,2,\ldots,n \text{ and } j= 1,2,\ldots,k \quad (3)$$

where $S_{jj} = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$ for explanatory variables, the product matrix of $Z'Z$ will produce correlation matrix of X as R.

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \ldots & r_{1k} \\ r_{12} & 1 & r_{23} & \ldots & r_{2k} \\ r_{13} & r_{23} & 1 & \ldots & r_{3k} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ r_{1k} & r_{2k} & r_{3k} & \ldots & 1 \end{bmatrix}$$

where $S_{ij} = \sum_{t=1}^{n}(X_{ui} - \bar{X}_i)(X_{uj} - \bar{X}_j)$, r_{ij} can be defined as:

$$r_{ij} = \frac{S_{ij}}{(S_{ii}S_{jj})^{1/2}}$$

Scaling reduces rounding errors, and also makes the condition indices comparable from one data set to another. However, if the intercept has interpretative value, then centering is not recommended (Montgomery 2001). Nonetheless, in some fields such as engineering and the physical sciences, the intercept has no physical interpretation. Examining the simple correlations r_{ij} (the off diagonal elements of the R matrix) is only helpful in detecting near–linear dependence between pairs of explanatory variables. Unfortunately, when more than two explanatory variables are involved in a near-linear dependence, there is no assurance that any of the pairwise correlations r_{ij}will be large (Montgomery 2001). Hence, generally, inspection of the r_{ij} is not sufficient for detecting anything more than pairwise multicollinearity.On the other hand the singular-value decomposition of $n\times p$ X matrix is defined as:

$$X = UDT' \tag{4}$$

where U is the $n\times p$ matrix whose columns are the eigenvectors associated with the p nonzero eigenvectors of XX'.T is the $p\times p$ matrix of eigenvectors of $X'X$. $UU'=I$, $TT'=I$, and D is a $p\times p$ diagonal matrix with nonnegative diagonal elements μ_j, $j=1,2,\ldots,p$ which are called singular-values of X matrix. $X'X = (UDT')'UDT' = TD^2T' = T\Lambda T'$ where Λ is a diagonal matrix whose diagonal elements are the eigen values of $X'X$ or $XX'(\lambda_1, \lambda_2, \ldots, \lambda_k)$ the squares of the μ_j. If we consider $R = X'X$ where R defined as the correlation matrix of unit normal scaled X, then T and Λ will be the eigen vectors and eigen values of R matrix.

Belsley et al. (1980) proposed an approach of diagnosing multicollinearity through the singular value decomposition of X matrix. They defined the Condition Indices (CI) of the X matrix as:

$$K_j = \frac{\lambda_{max}}{\lambda_j} \qquad j=1,\ldots,k \tag{5}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the eigen values of the matrix $X'X$ or correlation matrix of the unit normal scaled X. The largest value of K_j (called K) is defined as Condition Number (CN) of the matrix $X'X$. In addition, the square root of K is considered as the condition number of X matrix, which is preferably used as the condition number in Section 5. Condition numbers of X matrix between 10 and 30 imply moderate to strong multicollinearity while values more than 30 indicate severe multicollinearity.

The Ordinary Least Squares (OLS) estimate of regression model defined in equation (1) is given by:

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y \tag{6}$$

The covariance matrix of the estimates was defined as:

$$Var\left(\widehat{\beta}\right) = \sigma^2(X'X)^{-1} = \sigma^2 T\Lambda^{-1}\mathrm{T}' \tag{7}$$

and the variance of the j^{th} regression coefficient is:

$$Var\left(\hat{\beta}_j\right) = \sigma^2 \sum_{i=1}^{k} \frac{t_{ji}^2}{\lambda_i} = \sigma^2 VIF_j \qquad j = 1,2,\ldots,k \tag{8}$$

Hence, the j^{th} diagonal element of $T\Lambda^{-1}\mathrm{T}'$ is the j^{th} Variance Inflation Factor (VIF) that is one of the most practical multicollinearity diagnostic methods(Marquardt 1970).Obviously, when X matrix is unit length scaled, one or more small eigen values of correlation matrix (X) can inflate the variance of $\hat{\beta}_j$. The values of VIF between 5 and 10 indicate moderate to strong multicollinearity while the values of more than 10 diagnose the existence of severe multicollinearity in the data set. Belsley et al. (1980) also suggested using Variance Decomposition Proportions (VDP) as another measure of multicollinearity, which is defined and denoted as*:*

$$\pi_{ij} = \frac{t_{ji}^2/\lambda_i}{VIF_j} \qquad i,j = 1,2,\ldots,k \tag{9}$$

The elements of each column of π are just the proportions of the VIF or the variance of each $\hat{\beta}_j$ contributed by the i^{th} eigen values of the matrix $X'X$ or correlation matrix of the unit normal scaled X. Multicollinearity is indicated by the high proportion of the variance for two or more regression coefficients which is associated with one small eigen values of the matrix $X'X$ or correlation matrix of the unit normal scaled X , R. VDPs greater than 0.5 that corresponds to each of the large value of CN indicates the existence of one dependence in the column of X matrix. The explanatory variables that are involved in the multicollinear relationship in the X matrix column can be detected when the VDPs of the explanatory variables exceeds 0.5.

3. MINIMUM COVARIANCE DETERMINATION

The minimum covariance determinant (MCD) method of Rousseeuw (1985) is a highly robust estimator of the multivariate location and scatter (Andersen 2008; Daszykowski 2007; Rousseeuw and Leroy 1987; Wilcox 2005). Its objective is to find h observations (out of n) which covariance matrix has the lowest determinant. The MCD looks for the subset of data objects (h observations out of n) which sample covariance matrix has the smallest possible determinant. The MCD estimates of location and scatter are then the sample mean and sample covariance matrix

(multiplied by a consistency factor) of that optimal subset of h observations where h is usually chosen between ([n/2] + 1) and n ($[\frac{(n+k+1)}{2}]$ based on Splus 6 Robust Library User's Guide (Splus 6 robust library user's guide 2001). Although the raw minimum covariance determinant estimate has a high breakdown value, its statistical efficiency is low. Better finite-sample efficiency can be attained while retaining the high breakdown value by computing a one-step reweighted mean and a covariance estimate, with weights based on the MCD estimate. The one-step reweighted MCD estimates are a weighted mean and covariance, where regular observations are given weight one, but outliers (according to the initial MCD solution) are given weight zero. The one-step reweighted MCD estimators have the same breakdown value as the initial MCD estimators but with a much better finite-sample efficiency (Croux and Haesbroeck 1998; Lopuha¨a 1999; Rousseeuw and Van Driessen 1999). A one step reweighted MCD of location and scatter of X matrix, t_n^1 and C_n^1, are respectively defined as (Rousseeuw and Van Driessen 1999):

$$t_n^1 = (\sum_{i=1}^n w_i\, x_i)/(\sum_{i=1}^n w_i) \tag{10}$$

$$C_n^1 = \sum_{i=1}^n (w_i(x_i - t_n^1)(x_i - t_n^1)^T)/(\sum_{i=1}^n w_i - k) \tag{11}$$

By considering where

$$w_i = \begin{cases} 1 & if\ RMD_{MCD}(x_i) \le \sqrt{\chi^2_{(p,0.975)}} \\ 0 & otherwise \end{cases}$$

and

$$RMD_{MCD} = \sqrt{(X - \bar{X}_{MCD})'\sum\nolimits_{\mathrm{MCD}}{}^{-1}(X - \bar{X}_{MCD})} \qquad for \quad i = 1, \dots, n$$

Hereinafter, we will use the one-step reweighted MCD for the analysis and still refer them as MCD.

4. ROBUST MULTICOLLINEARITY DIAGNOSTIC MEASURES

There are several 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 2001). Kamruzzaman and Imon (2002) pointed out multiple high leverage points as new sources of multicollinearity where these points are observations that not only deviate from the same regression line as the other data in the data set, but also fall far from the majority of explanatory variables (Hocking and Pendelton 1983; Moller et al. 2005). All the high leverage points are not collinearity-influential observations and vice versa (Hadi 1988). High leverage points can reduce or increase collinearity according to their

position in the data. Classical multicollinearity diagnostic measures are not able to indicate collinearity pattern of a data in the presence of high leverage collinearity-influential observations especially when these points exist in a collinear data set (Bagheri et a;. 2010) .To avoid the distractive effect of these points on multicollinearity diagnostic measures, it is vital to utilize robust diagnostic measures (Bagheri et a;. 2010; Midi et al. 2010). It is worth mentioning that multicollinearity diagnostic measures depend on the eigen structure of non-scaled *X* matrix or correlation matrix (*X*) for scaled *X* matrix.

Midi et al. (2010) defined Robust Variance Inflation Factors (RVIF) by incorporating the robust coefficient determination based on Generalized M-estimator embedded with DRGP, namely as RVIF(GM (DRGP)) (For more details, one can refer to Midi et al. 2010) . They proved by numerical examples and simulation studies that this new proposed robust multicollinearity diagnostics is very powerful in indicating the source of multicollinearity for non-collinear data sets (Midi et al. 2010). Furthermore, RVF(GM(DRGP)) is resistant to the high leverage points and could successfully diagnose the multicollineariy pattern of collinear data sets (Bagheri and Habshah 2011). Besides the practical benefits of utilizing RVF(GM(DRGP)) in detecting the multicollinearity pattern of the data, it couldn't detect the variables which are in linear relationship with each other.

In what follows, we propose the robust multicollinearity diagnostic measure to solve the drawback of existing robust mulicollinearity diagnostic methods. To do so, for the *X* matrix which is scaled according to the equation (3), the MCD is incorporated in the development of the proposed measures. The Algorithm for the robust multicollinearity diagnostic measures based on MCD is summarized as follows;

Step 1: Scale the explanatory variables according to the Equation (3).

Step 2: Compute the product matrix of the scaled explanatory variables that is equal to the correlation matrix (*X*), *R*.

Step 3: Calculate robust estimate of the correlation matrix (*X*) based on MCD, denoted as R^{MCD}.

Step 4: Decompose $(R^{MCD})^{-1}$ as follows:

$$(R^{MCD})^{-1} = T^{MCD}(\Lambda^{MCD})^{-1}(T^{MCD})' \quad (12)$$

where $T^{mcd}$and $\Lambda^{mcd}$are the matrix of eigen vectors of R^{mcd} (where t_{ij}^{MCD} *i,j= 1,…,k* are the elements of T^{mcd} matrix) and diagonal matrix whose diagonal elements are the eigen values of R^{mcd} , ${\lambda_1}^{MCD}, {\lambda_2}^{MCD}, \dots, {\lambda_k}^{MCD}$, respectively.

Step 5: Finally the Robust Diagnostic Measures are formulated as follows:

- The Robust Condition Indices (RCI (MCD)) of the *X* matrix is defined as:

$$K_j^{MCD} = \frac{{\lambda_{max}}^{MCD}}{{\lambda_j}^{MCD}} \qquad j = 1, \dots, k \quad (13)$$

where ${\lambda_{max}}^{MCD}$ is the largest eigen values of R^{mcd} . The largest value of ${K_j}^{MCD}$ (called K^{MCD}) is defined as Robust Condition Number (RCN(MCD)) of the correlation matrix(X) based on MCD, that is R^{mcd}.

- The j^{th} VIF based on MCD is defined as:

$$VIF_j{}^{MCD} = \sum_{i=1}^{p} \frac{(t_{ji}^{MCD})^2}{{\lambda_i}^{MCD}} \qquad j = 1,2,\dots,k \qquad (14)$$

- The Robust Variance Decomposition Proportions (RVDP (MCD)) based on MCD also is formulated as*:*

$$\pi_{ij}{}^{\text{MCD}} = \frac{(t_{ji}^{MCD})^2/{\lambda_i}^{MCD}}{{VIF_j}^{MCD}} \qquad i,j = 1,2,\dots,k \qquad (15)$$

5. EXPERIMENTAL EVALUATION

5.1. BODY FAT DATA SET

In this section, the effect of high leverage points on a collinear data set, taken from Kutner et al. (2004) is investigated. Body Fat data set contains 20 observations with three explanatory variables, triceps skinfold thickness (X_1), thigh circumference (X_2) and mid arm circumference (X_3). This data set has multicollinearity problem (Kutner et al. 2004). In the collinear data set, the large magnitude of high leverage point in the same position of explanatory variables increase collinearity while in one or different positions of explanatory variables reduce linear dependency between them (Kutner et al. 2004).This data set is modified such that the first observation of the first explanatory variable is replaced with a large magnitude of high leverage point (equal to 300). Figure 1 demonstrates the original and the modified Body Fat data set. There is an obvious linear relationship between explanatory variables in the original data set (Figure 1 part (a)). However, after modifying the data set, the added high leverage point changed the collinearity pattern of the data (Figure 1 part (b)).

Table1 presents the classical and robust multicollinearity diagnostic measures for the Original Body Fat data set. According to this table, the Original Body Fat data set has severe multicollinearity, which can be concluded from the large value of the last CI or CN ($K_j = 53.3287 > 30$). Thus, there is one dependency in the column of X. Moreover, all the three values of π_{31}, π_{32} and π_{33} exceeded 0.5, indicating that all three explanatory variables are involved in a multicollinear relationship. The same results can be drawn from robust multicollinearity diagnostics based on the MCD.

The classical and robust multicollinearity diagnostic measures for Modified Body Fat

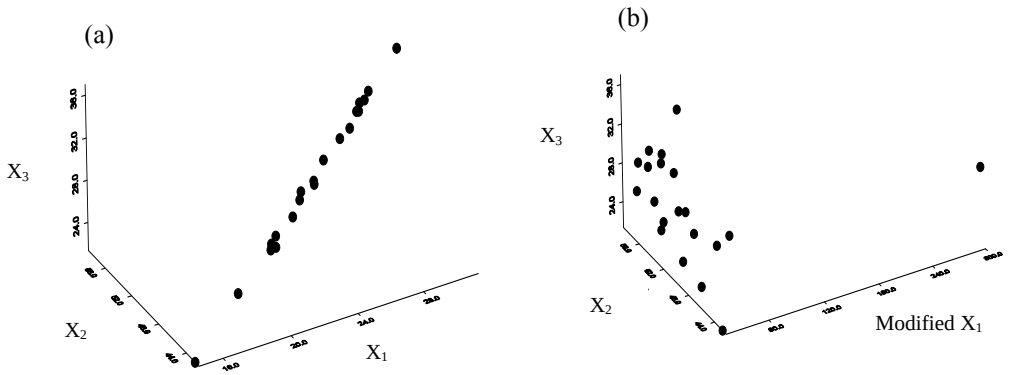

Figure.1 Original and modified Body fat data set

Table 1

Classical and robust multicollinearity diagnostic measures for Original Body fat data set

			VDP				RCI	RVDP(MCD)		
Ind.	λ	CI	X_1	X_2	X_3	λ^{MCD}	(MCD)	X_1	X_2	X_3
1	1.4375	1.0000	0.0003	0.0003	0.0006	1.4929	1.0000	0.0001	0.0001	0.0002
2	0.9658	1.4884	0.0000	0.0004	0.0082	0.8780	1.7003	0.0000	0.0002	0.0013
3	0.0270	53.3287	0.9997	0.9993	0.9912	0.0151	98.8294	0.9999	0.9997	0.9985

data set are illustrated in Table 2. It can be seen that none of the classical CI of X matrix were affected by the presence of the added high leverage point in this data set (CN < 30). However, robust multicollinearity diagnostic measures are resistant to the added high leverage points and indicate one collinear relationship between all explanatory variables in the data set. Table 3 exhibits the classical and robust VIF for the original and Modified Body Fat data set. Similar results of Table1 and Table 2 can be concluded from CVIF and RVIF. In the original data set, CVIF and RVIF (MCD) detected severe multicollinearity. After modifying the data set, CVIF failed to detect multicollinearity in the data while our new proposed RVIF is not affected by the added high leverage point and detected multicollinearity pattern of the data accurately.

5.2 MONTE CARLO SIMULATION STUDY

A simulation study was conducted to further assess the performance of our new proposed robust multicollinearity diagnostic measures based on MCD. Following the idea of Lawrence and Arthur (1990), three explanatory variables were generated as follows:

$$x_{ij} = (1-\rho^2)z_{ij} + \rho z_{i4} \quad for\ j = 1,2,3 \quad i = 1,\dots,n\,; \tag{16}$$

where the z_{ij} for $i = 1,\dots,n$ and $j = 1,\dots,4$ are independent standard normal random numbers. The value of ρ^2 represents the correlation between the two explanatory variables X_1 and X_2. In this study, ρ^2 was chosen to be equal to 0.98, which created high collinearity between these two explanatory variables. Thus in the data sets, only one dependence exists in the column of X, whereby two explanatory variables X_1 and X_2 are involved in this multicollinear relationship. Different percentages of high leverage points were added in the explanatory variables. The level of high leverage points (α) is varied from zero to 40%. The magnitude of high leverage points is varied from 20, 50,100 and 300. High leverage points in different positions of two collinear explanatory variables are generated in order to create collinearity-reducing observations. In doing so, the first 100 ($\frac{\alpha}{2}$) percent observations of X_1 and the last 100($\frac{\alpha}{2}$) percent observations of X_2 have been replaced by a certain magnitude of high leverage points. We ran 10000 simulations for moderate sample size equals to 100. The classical and robust multicollinearity diagnostic measures were then applied to the data. The results for the data without high leverage points are presented in Table 4. According to the results, the largest CI (CN) is greater than 30, which indicates that severe multicollinearity exists in the data sets. The VDP of X_1 and X_2 exceeded 0.5, which shows that these two explanatory variables caused multicollinearity in the simulated data sets. It is important to note that the same result can be drawn from RCN (MCD) and RVDP (MCD).

Table 2

Classical and robust multicollinearity diagnostic measures for Modified Body fat data set

			VDP				RCI	RVDP(MCD)		
Ind.	λ	CI	X_1	X_2	X_3	λ^{MCD}	(MCD)	X_1	X_2	X_3
1	1.1398	1.0000	0.3593	0.3149	0.0136	1.4929	1.0000	0.0001	0.0001	0.0002
2	1.0306	1.1060	0.0276	0.1202	0.7493	0.8780	1.7003	0.0000	0.0002	0.0013
3	0.7992	1.4263	0.6132	0.5649	0.2371	0.0151	98.8294	0.9999	0.9997	0.9985

Table 5 presents the classical and robust multicollinearity diagnostic measures for *n*=100, different percentages of high leverage points, and at of 20 and 300 magnitude

Table 3

Classical and robust VIF for original and modified Body fat data set

	Original data set		Modified data set	
index	CVIF	RVIF(MCD)	CVIF	RVIF(MCD)
1	708.8429	2718.3973	1.1266	2718.3973
2	564.3434	1761.5475	1.1141	1761.5475
3	104.6060	496.1453	1.0363	496.1453

Table 4

Simulated classical and robust multicollinearity diagnostic measures for *n*=100 and without high leverage points

			VDP				RCI	RVDP(MCD)		
Ind.	λ	CI	X_1	X_2	X_3	λ^{MCD}	(MCD)	X_1	X_2	X_3
1	1.4204	1.0000	0.0008	0.0008	0.0082	1.4849	1.0000	0.0004	0.0004	0.0356
2	0.9903	1.4343	0.0000	0.0000	0.9815	0.8783	1.6907	0.0003	0.0003	0.7636
3	0.0400	**35.5100**	**0.9992**	**0.9992**	0.0103	0.0265	**56.0340**	**0.9994**	**0.9994**	0.2007

of high leverage points. The results of other magnitudes of high leverage points are consistent and are not included due to space limitation. It is worth mentioning that the values of CI and RCI (MCD) are computed based on the values of λ and λ^{MCD}, respectively. As soon as the small magnitude and percentages of high leverage points were added to the simulated data sets, the classical multicollinearity diagnostic measures such as the CI failed to detect the presence of multicollinearity in the data set. Since none of the CIs exceeded their cutoff points, the results of VDP were not checked because it was not conclusive as to which explanatory variables were involved in the collinearity relationship. However, the values of the last RCI (MCD)s, RCN (MCD)s for different percentages and magnitudes of high leverage points, are greater than the cutoff point 30, which indicates that they correctly specify one linear relationship in the data set. On the other hand, the RVDP (MCD) s in all different situations recognized the explanatory variables which are involved in this collinear relationship.

Figure 2 part (a) and part (b) illustrate the maximum classical VIF and robust VIF based on MCD for different magnitudes and percentages of high leverage points. This Figure part (a) clearly shows the effect of adding high leverage points in collinear simulated data sets. When none of the high leverage points are present in the data sets.

Table 5

Simulated classical and robust multicollinearity diagnostic measures for n=100 and different percentage and magnitude of high leverage points

					VDP				RCI	RVDP(MCD)		
MC	α	I	λ	CI	X_1	X_2	X_3	λ^{MCD}	(MCD)	X_1	X_2	X_3
20	10	1	1.0655	1.0000	0.2623	0.2589	0.3458	1.4762	1.0000	0.0004	0.0004	0.0351
		2	0.9995	1.0661	0.4002	0.4061	0.1875	0.8952	1.649063	0.0002	0.0002	0.7932
		3	0.9294	1.1465	0.3375	0.3351	0.4668	0.0276	**53.49326**	**0.9994**	**0.9994**	0.1717
	20	1	1.0814	1.0000	0.3174	0.3154	0.2029	1.4681	1.0000	0.0004	0.0004	0.0347
		2	0.9974	1.0842	0.2453	0.2502	0.4966	0.9111	1.6113	0.0002	0.0002	0.8282
		3	0.9130	1.1844	0.4373	0.4344	0.3004	0.0292	**50.3217**	**0.9994**	**0.9994**	0.1371
	30	1	1.1052	1.0000	0.3366	0.3355	0.1153	1.4571	1.0000	0.0005	0.0005	0.0322
		2	0.9948	1.1109	0.1405	0.1434	0.7072	0.9311	1.565011	0.0001	0.0001	0.8654
		3	0.8872	1.2457	0.5229	0.5211	0.1774	0.0311	**46.82461**	**0.9994**	**0.9994**	0.1024
	40	1	1.1346	1.0000	0.3249	0.3260	0.0738	1.4422	1.0000	0.0006	0.0006	0.0261
		2	0.9937	1.1418	0.0886	0.0853	0.8138	0.9565	1.507876	0.0001	0.0001	0.9133
		3	0.8508	1.3335	0.5864	0.5887	0.1124	0.0345	**41.84453**	**0.9993**	**0.9993**	0.0605
300	10	1	1.0667	1.0000	0.2651	0.2650	0.3345	1.4774	1.0000	0.0004	0.0004	0.0351
		2	0.9992	1.0675	0.3922	0.3920	0.2098	0.8933	1.6539	0.0002	0.0002	0.7903
		3	0.9282	1.1491	0.3427	0.3430	0.4557	0.0276	**53.5531**	**0.9994**	**0.9994**	0.1746
	20	1	1.0819	1.0000	0.3212	0.3169	0.1967	1.4678	1.0000	0.0004	0.0004	0.0343
		2	0.9971	1.0851	0.2361	0.2456	0.5111	0.9115	1.6104	0.0002	0.0002	0.8270
		3	0.9128	1.1852	0.4427	0.4375	0.2922	0.0291	**50.4219**	**0.9994**	**0.9994**	0.1387
	30	1	1.1056	1.0000	0.3355	0.3368	0.1141	1.4561	1.0000	0.0005	0.0005	0.0318
		2	0.9951	1.1110	0.1419	0.1389	0.7096	0.9330	1.5607	0.0001	0.0001	0.8667
		3	0.8864	1.2473	0.5226	0.5243	0.1762	0.0311	**46.8579**	**0.9994**	**0.9994**	0.1015
	40	1	1.1351	1.0000	0.3246	0.3255	0.0737	1.4416	1.0000	0.0006	0.0006	0.0257
		2	0.9935	1.142546	0.0876	0.0853	0.8153	0.9575	1.505562	0.0001	0.0001	0.9145
		3	0.8504	1.334718	0.5879	0.5892	0.1111	0.0346	**41.69594**	**0.9993**	**0.9993**	0.0598

#MC: magnitude of high leverage points and #I :Index.

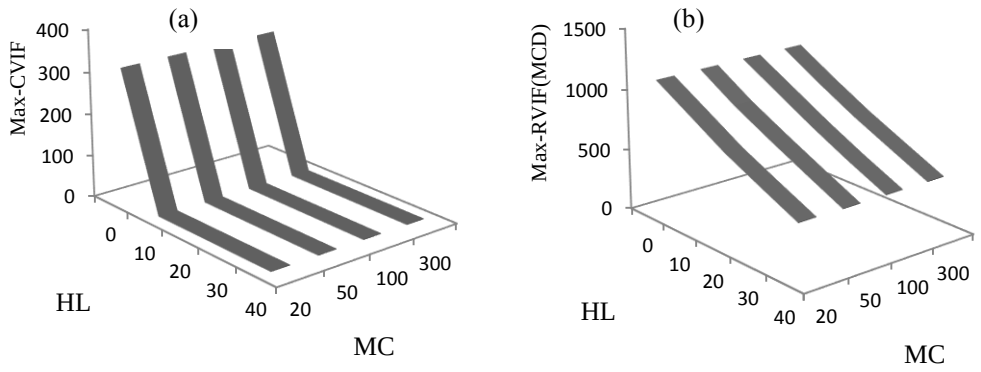

Figure 2. Maximum values of CVIF or RVIF (MCD) for different percentage and magnitude of high leverage points and *n*=100

the maximum CVIF for all different magnitudes of high leverage points exceed the cutoff point, which reflects that without high leverage points, the data sets are collinear. By increasing the magnitude of the high leverage points, the CVIF fails to detect collinearity in the data. Figure 2 part (b) reveals that if RVIF (MCD) is used as an indicator of multicollinearity diagnostic, it can diagnose correctly the collinearity of the data sets. However, by increasing the percentage of high leverage points decreases the values of maximum RVIF (MCD).

5. CONCLUSION

High leverage points that are the outliers in the *X*-direction may change the multicollinearity pattern of the data set. The classical multicollinearity diagnostic measures in the presence of these points in a collinear data set may fail to detect the multicollinearity in the data set. Moreover, the existing robust multicollinearity diagnostics tools are not able to detect the collinear explanatory variables. In this paper, we proposed robust multicollinearity diagnostic measures based on a one-step reweighted Minimum Covariance Determinate (MCD). The performance of our robust multicollinearity diagnostic measures is evaluated by a well-known collinear data set and simulation study. The results of the study signify that all the classical measures failed to detect the multicollinearity pattern of the data whereas our proposed measures were very successful in detecting the multicollinearity pattern of the data and the collinear explanatory variables.

REFERENCES

[1] **Andersen, R.(2008),** ***Modern Methods for Robust Regression***; Sara Miller McCune :SAGE;

[2] **Bagheri, A. & Habshah, M. (2011),** ***On the Performance of Robust Variance Inflation Factors***. *International Journal of Agricultural Statistical Science*. **7**(1), 31-45;

[3] **Bagheri, A., Habshah, M. & Imon, A.H.M.R. (2011),** ***A Novel Collinearity-Influential Observation Diagnostic Measure Based on a Group Deletion Approach***. *Journal of Communication and Statistics, Simulation and Computation*, 41:8,1379-1396;

[4] **Bagheri, A., Midi, Habshah & Imon, AHMR (2010),** ***The Effect of Collinearity- influential Observations on Collinear Data Sets; A Monte Carlo Simulation Study".*** *Journal of Applied Science*, 10, 2086-2093;

[5] **Belsley, DA , Kuh, E & Welsch, RE.(1980),** ***Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*** ; New York: *Wiley*.

[6] **Croux, C & Haesbroeck, G. (1998),** ***Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator*** *; Journal of Multivariate Analysis*.71,161-190;

[7] **Daszykowski, M, Kaczmarek, K, Vander Heyden, Y & Walczak, B. (2007),** ***Robust Statistics in Data Analysis - A Review Basic Concepts***; *Chemometrics and Intelligent Laboratory*, 85, 203–219;

[8] **Hadi, AS. (1988),** ***Diagnosing Collineariy-influential Observations*****.** *Computational Statistic Data Analysis*.7, 143-159;

[9] **Hocking, RR & Pendelton, OJ.(1983),** ***The Regression Dilemma***. *Communications in Statistics Theory and Methods*.12, 497-527;

[10] **Kamruzzaman, MD & Imon, AHMR.(2002),** ***High Leverage Point: Another Source of Multicollinearity*** . *Pakistanian Journal of Statistics*, 18, 435-448;

[11] **Kutner, MH, Nachtsheim, CJ & Neter, J.(2004),** ***Applied Linear Regression Models***; 4th Edition, New York:*MacGRAW-Hil;*

[12] **Lawrence, K & Arthur, JL. (1990),** ***Robust Regression; Analysis and Applications*** ; Marcel Dekker: INC.

[13] **Lopuha¨a, HP, (1999),** ***Asymptotics of Reweighted Estimators of Multivariate Location and Scatter*** . *Annals of Statistics*.27,1638–166;

[14] **Marquardt, DW. (1970),** ***Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation*** *; Technometrics*. 12, 591-612;

[15] **Midi, H, Bagheri, A. & Imon AHMR.(2010),** ***The Application of Robust Multicollinearity Diagnostics Method Based on Robust Coefficient Determination to a Non-collinear Data Set***. *Journal of Applied Science*. 10(8), 611-619;

[16] **Moller, SF, Frese, JV & Bro, R.(2005),** ***Robust Methods for Multivariate Data Analysis***. *Journal of Chemometrics*.19, 549-563;

[17] **Montgomery, DC, Peck, EA & Vining, GG.(2001),** ***Introduction to Linear Regression Analysis****;* 3th Edition, New York: *Wiley*.

[18] **Rousseeuw, PJ.(1985),** ***Multivariate Estimation with High Breakdown Point***. *Mathematical Statistics with Application*. B, 283-297;

[19] **Rousseeuw, PJ & Leroy, AM. (1987),** ***Robust Regression and Outlier Detection*** *;* New York: *John Willy;*

[20] **Rousseeuw, PJ & Van Driessen, K.(1999),** ***A Fast Algorithm for the Minimum Covariance Determinant Estimator*** ;*Technometrics*.41,212–223;

[21] Splus 6 robust library user's guide. (2001), insightful corporation, seattle, Wa.

[22] **Wilcox, RR.(2005),** ***Introduction to Robust Estimation and Hypothesis Testing****;* 2nd Edition, The United States of America :*Elsevier Academic Press*.