Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 531607, 16 pages
doi:10.1155/2012/531607

Research Article

On the Performance of the Measure for Diagnosing Multiple High Leverage Collinearity-Reducing Observations

Arezoo Bagheri[1] and Habshah Midi[1,2]

[1] *Laboratory of Computational Statistics and Operations Research, Institute for Mathematical Research, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*
[2] *Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia*

Correspondence should be addressed to Habshah Midi, habshahmidi@gmail.com

Received 2 August 2012; Revised 9 December 2012; Accepted 9 December 2012

Academic Editor: Stefano Lenci

Copyright © 2012 A. Bagheri and H. Midi. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

There is strong evidence indicating that the existing measures which are designed to detect a single high leverage collinearity-reducing observation are not effective in the presence of multiple high leverage collinearity-reducing observations. In this paper, we propose a cutoff point for a newly developed high leverage collinearity-influential measure $\delta_i^{(D)}$ and two existing measures (δ_i and l_i) to identify high leverage collinearity-reducing observations, the high leverage points which hide multicollinearity in a data set. It is important to detect these observations as they are responsible for the misleading inferences about the fitting of the regression model. The merit of our proposed measure and cutoff point in detecting high leverage collinearity-reducing observations is investigated by using engineering data and Monte Carlo simulations.

1. Introduction

High leverage points are the observations that fall far from the majority of explanatory variables in the data set (see [1–4]). It is now evident that high leverage point is another prime source of multicollinearity; a near-linear dependency of two or more explanatory variables [2]. Hadi [5] pointed out that this source of multicollinearity is a special case of collinearity-influential observations; the observations which might induce or disrupt the multicollinearity pattern of a data. High leverage points that induce multicollinearity are referred as high leverage collinearity-enhancing observations while those that reduce multicollinearity in their presence are called high leverage collinearity-reducing observations

[6–10]. Collinearity-influential observations are usually points with high leverages, though all high leverage points are not necessarily collinearity-influential observations [5].

It is very important to detect collinearity-influential observations because they are responsible for misleading conclusion about the fitting of a regression model, which gives wrong sign problem of regression coefficients and produces large variances to the regression estimates. Not many studies have been conducted in the literature on collinearity-influential measures and we will discuss these methods in Section 2. Nonetheless most of the existing methods are not successful in the detection of multiple high leverage collinearity-influential observations although their performances are considered good for the detection of a single observation. Moreover these measures do not have specific cutoff points to indicate the existence of collinearity-influential observations [10]. These shortcomings motivated us to propose a new detection measure in such situation. Notably, the proposed measure is based on the Diagnostic Robust Generalized Potential (DRGP) method developed by Habshah et al. [11] and will be presented in Section 3. Section 4 exhibits the development of the collinearity-influential observations that can be classified as high leverage collinearity-enhancing or collinearity-reducing observations. Bagheri et al. [10] presented numerical examples and a simulation study to propose a novel high leverage collinearity-influential measure and a cutoff point for the detection of high leverage collinearity-enhancing observations. The authors also recommended cutoff points for collinearity-influential measures introduced by Hadi [5] and Sengupta and Bhimasankaram [12]. It is also important to identify high leverage collinearity-reducing observations. However, these observations are more difficult to diagnose because they hide the effect of multicollinearity in the classical analysis. Following Hadi [13], Imon [14], and Habshah et al. [11], in Section 5, we propose a cutoff point for Bagheri's et al. [10], Hadi [5], and Sengupta and Bhimasankaram [12]'s measures to identify high leverage collinearity-reducing observations. A numerical example and simulation study are performed in Sections 6 and 7, respectively, to evaluate the performance of our proposed measure ($\delta_i^{(D)}$) and compare its performance with Hadi [5] and Sengupta and Bhimasankaram [12]'s measures (δ_i and l_i). Conclusion of the study will be presented in Section 8.

2. Collinearity-Influential Measures

Let consider a multiple linear regression model as follows:

$$Y = X\beta + \varepsilon, \tag{2.1}$$

where Y is an $(n \times 1)$ vector of response or dependent variable, X is an $(n \times p)$ matrix of predictors $(n > p)$, β is a $(p \times 1)$ vector of unknown finite parameters to be estimated and ε is an $(n \times 1)$ vector of random errors. We let X_j denote the jth column of the X matrix; therefore, $X = [X_1, X_2, \ldots, X_p]$. Furthermore, multicollinearity is defined in terms of the linear dependence of the columns of X.

Belsley et al. [15] proposed the singular-value decomposition of $(n \times p)$ X matrix for diagnosing multicollinearity as follows:

$$X = UDV^T, \tag{2.2}$$

where U is the $(n \times p)$ matrix in which the columns that are associated with the p nonzero eigenvalue of (X^TX) is $(n \times p)$, V (the matrix of eigenvectors of X^TX) is $(p \times p)$, $U^TU = I$, $V^TV = I$, and D is a $(p \times p)$ diagonal matrix with nonnegative diagonal elements, k_j, $j = 1, 2, \ldots, p$, which is called singular-values of X. Condition number of X matrix denoted as CN is another multicollinearity diagnostic measures which is obtained by first computing the Condition Index (CI) of the X matrix and is defined as

$$k_j = \frac{\lambda_{\max}}{\lambda_j}, \quad j = 1, 2, \ldots, p, \tag{2.3}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the singular values of the X matrix. The CN corresponds to the largest values of k_j. To make the condition indices comparable from one data set to another, the independent variables should first be scaled to have the same length. Scaling the independent variables prevents the eigen analysis to be dependent on the variables' units of measurements. Belsley [16] stated that CN of X matrix between 10 to 30 indicates moderate to strong multicollinearity, while a value of more than 30 reflects severe multicollinearity.

Hadi [5] noted that most collinearity-influential observations are points with high leverages, but not all high leverage points are collinearity-influential observations. He defined a measure for the influence of the ith row of X matrix on the condition index denoted as δ_i,

$$\delta_i = \frac{k_{(i)} - k}{k}, \quad i = 1, 2, \ldots, n, \tag{2.4}$$

where $k_{(i)}$ is computed by the eigenvalue of $X_{(i)}$ and when the ith row of X matrix has been deleted. Due to the lack of symmetry of Hadi's measure, Sengupta and Bhimasankaram [12] proposed a collinearity-influential measure for each row of observations, defined as

$$l_i = \log\left(\frac{k_{(i)}}{k}\right), \quad i = 1, 2, \ldots, n. \tag{2.5}$$

Unfortunately, they did not propose practical cutoff points for δ_i and l_i and only mentioned the conditions for collinearity-enhancing and collinearity-reducing observations. To fill the gap, Bagheri et al. [10] suggested a cutoff point for δ_i and l_i for detecting collinearity-enhancing observations as

$$\text{cut(CEO)} = \text{Median}(\theta_i) - 3\text{MAD}(\theta_i), \quad i = 1, 2, \ldots, n, \tag{2.6}$$

where cut(CEO) is the Collinearity-Influential Measure cutoff point for the identification of collinearity-enhancing observations whereby θ_i can be δ_i or l_i. $|\theta_i| \geq |\text{cut(CEO)}|$ for $\theta_i < 0$ is an indicator that the ith observation is a collinearity-enhancing observation.

3. Diagnostics Robust Generalised Potential for Identification of High Leverage Points

The ith diagonal elements of the hat matrix, $W = X(X^TX)^{-1}X^T$, is a traditionally used measure for detecting high leverage points and is defined as

$$w_{ii} = x_i^T\left(X^TX\right)^{-1}x_i, \quad i = 1, 2, \ldots, n. \tag{3.1}$$

Hoaglin and Welsch [17] suggested twice-the-mean-rule $(2(p+1)/n)$ cutoff points for the hat matrix. Hadi [13] pointed out that the leverage diagnostics may not be successful to identify high leverage points and introduced a single-case-deleted measure, known as potential, and is defined as

$$p_{ii} = x_i^T\left(X_{(i)}^TX_{(i)}\right)^{-1}x_i, \quad i = 1, 2, \ldots, n \tag{3.2}$$

or

$$p_{ii} = \frac{w_{ii}}{1 - w_{ii}}, \quad i = 1, 2, \ldots, n, \tag{3.3}$$

where $X_{(i)}$ is the data matrix X with the ith row deleted. Imon [14] pointed that potentials may be very successful in the identification of a single high leverage point, but they fail to identify multiple high leverage points. To rectify this problem, Imon [14] proposed a group deletion version of potentials (GP), known as generalized potentials. Prior to defining the GP, Imon [14] partitioned the data into a set of "good" cases "remaining" in the analysis and a set of "bad" cases "deleted" from the analysis which were denoted as R and D. Nonetheless, Imon's measure has drawbacks which are due to the inefficient procedure that he used for the determination of the initial deletion set D. To overcome this shortcoming, Habshah et al. [11] proposed the diagnostic robust generalized potential (DRGP) where the suspected cases (bad cases) were identified by Robust Mahalanobis Distance (RMD), based on the Minimum Volume Ellipsoid (MVE). Rousseeuw [18] defined RMD based on MVE as follows:

$$\mathrm{RMD}_i = \sqrt{(X - T_R(X))^T C_R(X)^{-1}(X - T_R(X))} \quad \text{for } i = 1, 2, \ldots, n, \tag{3.4}$$

where $T_R(X)$ and $C_R(X)$ are robust locations and shape estimates of the MVE, respectively. In the second step of DRGP (MVE), the GPs are computed based on the set of D and R obtained from RMD (MVE). The low leverage points (if any) are put back into the estimation data set after inspecting the GP proposed by Imon [14] which are defined as follows:

$$p_{ii}^* = \begin{cases} w_{ii}^{(-D)} & \text{for } i \in D, \\ \dfrac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} & \text{for } i \in R, \end{cases} \tag{3.5}$$

where $w_{ii}^{(-D)} = X_i^T (X_R^T X_R)^{-1} X_i$. He suggested the cutoff point of p_{ii}^* as

$$p_{ii}^* > \text{Median}(p_{ii}^*) + c\text{MAD}(p_{ii}^*), \tag{3.6}$$

where c can be taken as a constant value of 2 or 3.

The DRGP (MVE) have been proven to be very effective in the identification of multiple high leverage points.

4. The New Proposed High Leverage Collinearity-Influential Observations Measures

As already mentioned in the preceding section, the main reason of developing a new measure of high leverage collinearity-influential measure is due to the fact that the commonly used measures failed to detect multiple high leverage collinearity-influential observations. In addition, not many papers related to this measure have been published in the literatures. It is important mentioning that the collinearity-influential measure which were proposed by Hadi [5] and Sengupta and Bhimasankaram [12] are related to the Hadi's single-case-deleted leverage measure [13]. Since the robust generalized potentials that was developed by Habshah et al. [11] was very successful in the identification of multiple high leverage points compared to other widely used methods, Bagheri et al. [10] utilized a similar approach in developing multiple High Leverage Collinearity-Influential Measure (HLCIM). The proposed measure is formulated based on Sengupta and Bhimasankaram [12]'s measure with slight modification whereby almost similar approach of DRGP (MVE) [11] was adapted. Hence it is referred as HLCIM (DRGP) and denoted as $\delta_i^{(D)}$. This new measure is defined as follows:

$$\delta_i^{(D)} = \begin{cases} \log\left(\dfrac{k_{(D)}}{k_{(D-i)}}\right) & \text{if } i \in D,\ \#\{D\} \neq 1, \\ \log\left(\dfrac{k_{(i)}}{k}\right) & \text{if } \#\{D\} = 1,\ D = i,\ i = 1, 2, \ldots, n, \\ \log\left(\dfrac{k_{(D+i)}}{k_{(D)}}\right) & \text{if } i \in R, \end{cases} \tag{4.1}$$

where D is the suspected group of multiple high leverage collinearity-influential observations diagnosed by DRGP(MVE), p_{ii}^*, $\#\{D\}$ is the number of elements in D group, and R is the remaining good observations. As such, following Habshah et al. [11] approach, three conditions should be considered in defining $\delta_i^{(D)}$. Bagheri et al. [10] summarized the algorithm of HLCIM (DRGP) in three steps as follows.

Step 1. Calculate DRGP (MVE), p_{ii}^*, for $i = 1, 2, \ldots, n$. Form D as a high leverage collinearity-influential suspected group whereby its members consist of observations which correspond to p_{ii}^* that exceed the median(p_{ii}^*) + 3MAD(p_{ii}^*). Obviously the rest of the observations belong to R, the remaining group.

Step 2. Compute high leverage collinearity-influential values, $\delta_i^{(D)}$, as follows.

(i) If only a single member in the D group, the size of R is $(n-1)$, and $D = i$, calculate $\log(k_{(i)}/k)$ where $k_{(i)}$ indicates the condition number of the X matrix without the ith high leverage points. In this way, $\delta_i^{(D)} = l_i$.

(ii) If more than one member in the D group, calculate $\log(k_{(D)}/k_{(D-i)})$ where $k_{(D-i)}$ indicates the condition number of the X matrix without the entire D group minus the ith high leverage points, where i belongs to the suspected D group.

(iii) For any observation in the R group, compute $\log(k_{(D+i)}/k_{(D)})$ where $k_{(D+i)}$ refers to the condition number of the X matrix without the entire group of D high leverage points plus the ith additional observation of the remaining group.

Step 3. If any $\delta_i^{(D)}$ values for $i = 1, 2, \ldots, n$ does not exceed the cutoff points in (2.6), put back the ith observation to the R group. Otherwise, D group is the high leverage collinearity-enhancing observations.

Bagheri et al. [10] only defined the cutoff point for θ_i to indicate high leverage collinearity-enhancing observations and they did not suggest cutoff point for collinearity-reducing observations. The authors considered θ_i to be high leverage collinearity-enhancing observations if θ_i is less than the cutoff points; that is $\text{median}(\theta_i) - 3\text{mad}(\theta_i)$ for $\theta_i < 0$, where c is a chosen value 3 and θ_i may be $\delta_i^{(D)}$, δ_i or l_i.

Since high leverage collinearity-reducing observations are also responsible for the misleading inferential statements, it is very crucial to detect their presence. In the following section, we propose a cutoff point for identifying high leverage collinearity-reducing observations.

It is important mentioning that not all $\delta_i^{(D)}$ which exceed the cutoff point are high leverage points. This is true for the situation when $\delta_i^{(D)}$ exceeds the cutoff point but belongs to the remaining group, $i \in R$. In this situation, the observation is considered as collinearity-influential observations since they are not high leverage points.

5. The New Proposed Cutoff Point for HLCIM (DRGP)

Hadi [5] and Sengupta and Bhimasankaram [12] mentioned that a large positive value of their collinearity-influential measures, δ_i and l_i, respectively, indicates that the ith observation is a collinearity-reducing observation. However, they did not suggest any cutoff points to indicate which observations are collinearity-enhancing and which are collinearity-reducing. Bagheri et al. [10] proposed a nonparametric cutoff point for high leverage collinearity-enhancing observations. Their work has inspired us to investigate high leverage collinearity-reducing observations among the observations that correspond to positive values of high leverage collinearity-influential measures. Figure 1 presents the normal distribution plot of θ_i. Based on this figure, any value that exceeds $\text{median}(\theta_i) + 3\text{MAD}(\theta_i)$ can be utilized as a cutoff point for θ_i. Hence, we propose the following cutoff point:

$$\text{cut(CRO)} = \text{Median}(\theta_i) + 3\text{MAD}(\theta_i), \tag{5.1}$$

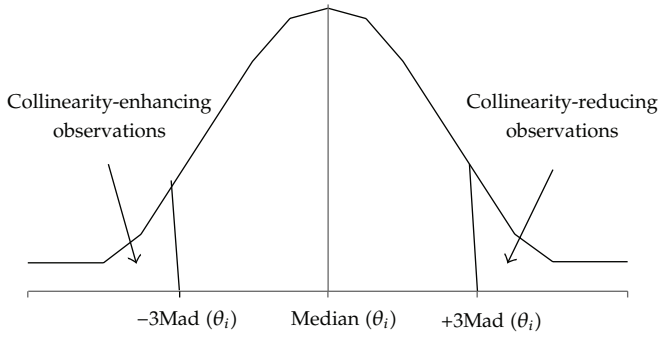


Figure 1: Normal distribution plot of high leverage collinearity-influential measure.

where cut(CRO) is the Collinearity-Influential Measure cutoff point for Collinearity-Reducing Observations. θ_i can be $\delta_i^{(D)}$, δ_i or l_i. $\theta_i \geq$ cut(CRO) for $\theta_i > 0$ is an indicator that the ith observation is a collinearity-reducing observation.

6. A Numerical Example

A numerical example is presented to compare the performance of the newly proposed measure ($\delta_i^{(D)}$) with the existing measures δ_i and l_i. An engineering data taken from Montgomery et al. [19] is used in this study. It represents the relationship between thrust of a jet-turbine engine (y) and six independent variables. The independent variables are primary speed of rotation (X_1), secondary speed of rotation (X_2), fuel flow rate (X_3), pressure (X_4), exhaust temperature (X_5), and ambient temperature at time of test (X_6). It is important mentioning that, the explanatory variables of this data are scaled before analysis in order to prevent the condition number to be dominated by large measurement units of some explanatory variables. Prior to analysis of this data, the explanatory variables have been scaled following Stewart's [20] scaling method as

$$x'_{ij} = \frac{x_{ij}}{\|X_j\|}, \quad i = 1, \ldots, p, \ j = 1, \ldots, n. \tag{6.1}$$

There are other alternative scaling methods which can be found in Montgomery et al. [1], Stewart [20], and Hadi [5].

The matrix plot in Figure 2 and the collinearity diagnostics presented in Table 1 suggest that this data set has severe multicollinearity problem (CN = 47.78). We would like to diagnose whether high leverage points are the cause of this problem. As such, it is necessary to detect the presence of high leverage points in this data set.

The index plot of DRGP (MVE) presented in Figure 3 suggests that observations 6 and 20 are high leverage points. By deleting these two observations from the data set, CN increases to 52.09. It seems that these two high leverages are collinearity-reducing observations.

The effect of these two high leverage points on collinearity pattern of the data is further investigated by applying $\delta_i^{(D)}$, δ_i and l_i with their respective new cutoff point introduced in (5.1) for detecting high leverage collinearity-reducing observations. Figure 4 illustrates

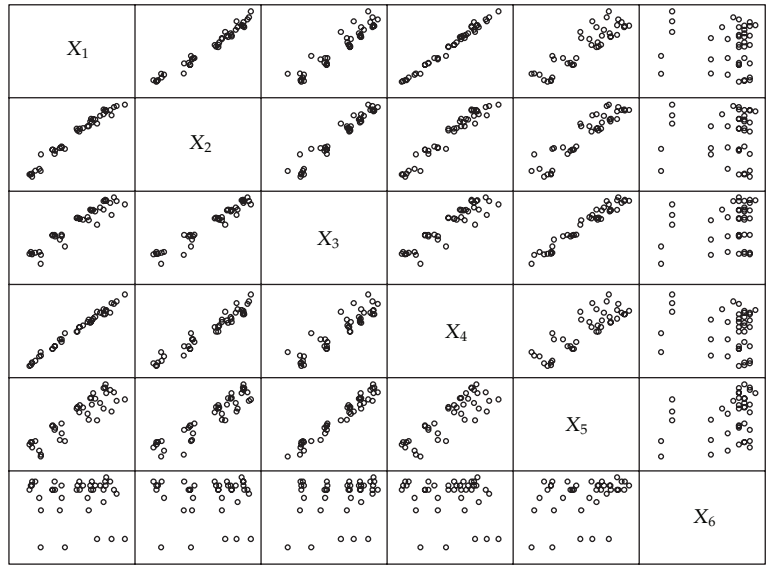


Figure 2: Matrix plot of jet turbine engine data set.

Table 1: Collinearity diagnostics of jet turbine engine data set.

Diagnostics	1	2	3	4	5	6
Pearson correlation coefficient	$r_{12} = 0.98$	$r_{13} = 0.95$	$r_{14} = 0.99$	$r_{15} = 0.89$	$r_{16} = -0.07$	$r_{23} = 0.97$
	$r_{24} = 0.97$	$r_{25} = 0.93$	$r_{26} = 0.02$	$r_{34} = 0.92$	$r_{35} = 0.98$	$r_{36} = 0.22$
	$r_{45} = 0.85$	$r_{46} = -0.15$	$r_{56} = 0.30$			
VIF >5	289.11	71.83	168.05	219.97	32.41	8.48
Condition index of X matrix >10	47.78	32.22	18.65	10.17	2.04	1.00

the index plot of these measures. According to this plot, all these three measures have indicated that observations 6 and 20 as high leverage collinearity-reducing observations. Nevertheless, besides observations 6 and 20, they detect a few more observations as collinearity-reducing observations. It is interesting to note that none of the observations are detected as high leverage collinearity-enhancing observations or collinearity-enhancing observations.

It is worth mentioning that we do not have any information about the source of the two existing high leverage collinearity-reducing observations (cases 6 and 20). Therefore, we cannot control the magnitude and the number of added high leverages points to the data in order to study the effectiveness of our proposed measures. In this respect, we have modified this data set in two different patterns following [7]. Habshah et al. [7] indicated that in the collinear data set, when high leverages exist in just one explanatory variable or in different positions of two explanatory variables; these leverages will be collinearity-reducing observations. Thus, the first pattern is when we replaced observations 5, 6, 19, and 20 of X_2 with a fixed large value of 50000. The second pattern is created by replacing the large value of 50000 to X_2 for observations 5, 6 and observations 19, 20 of X_3.

The DRGP (MVE) index plot for Figure 5 reveals that observations 5, 6, 19, 20 are detected as high leverage points for modified jet turbine engine data set.

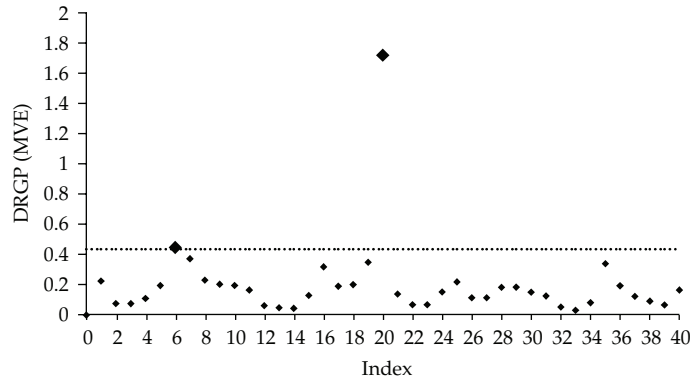


Figure 3: DRGP(MVE) index plot of jet turbine engine data set.

Figures 6 and 7 present the index plot of $\delta_i^{(D)}$, δ_i and l_i for the first and the second pattern of the modified jet turbine engine data set. The results of $\delta_i^{(D)}$ in these figures agree reasonably well with Bagheri's et al. [10] findings that when high leverage points exist in just one explanatory variable (first pattern) or in different positions of two explanatory variables (second pattern) in collinear data sets, these observations are referred as collinearity-reducing observations. For both patterns, $\delta_i^{(D)}$ correctly identified that observations 5, 6, 19, and 20 are high leverage collinearity-reducing observations. However, for the first pattern, both δ_i and l_i are not successful in detecting all of observations; 5, 6, 19, and 20 as high leverage collinearity-reducing observations. In the first pattern, they only correctly detected observations 19 and 20 as high leverage collinearity-reducing observations. However, none of the added high leverage collinearity-reducing observations can be detected by these two measures in the second pattern. It is important to note that for the first and the second patterns, the values of δ_i and l_i for the observations 5 and 6, and observation 19, respectively are becoming negative. This indicates that for both patterns, δ_i and l_i have wrongly indicated these observations as suspected high leverage collinearity-enhancing observations.

7. Monte Carlo Simulation Study

In this section, we report a Monte Carlo simulation study that is designed to assess the performance of our new proposed measure $\delta_i^{(D)}$ in detecting multiple high leverage collinearity-reducing observations and to compare its performance with two commonly used measures (δ_i and l_i). Following Lawrence and Arthur [21], simulated data sets with three independent regressors were generated as follows:

$$x_{ij} = \left(1 - \rho^2\right) z_{ij} + \rho z_{i4}, \quad i = 1, \ldots, n;\ j = 1, \ldots, 3, \tag{7.1}$$

where the z_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, 3$ are Uniform (0, 1). The value of ρ^2 which represents the correlation between the two explanatory variables are chosen to be equal to 0.95. This amount of correlation causes high multicollinearity between explanatory variables. Different percentage of high leverage points are considered in this study. The level of high leverage points varied from $\alpha = 0.10, 0.20, 0.30$. Different sample sizes from $n = 20, 40, 60, 100$, and

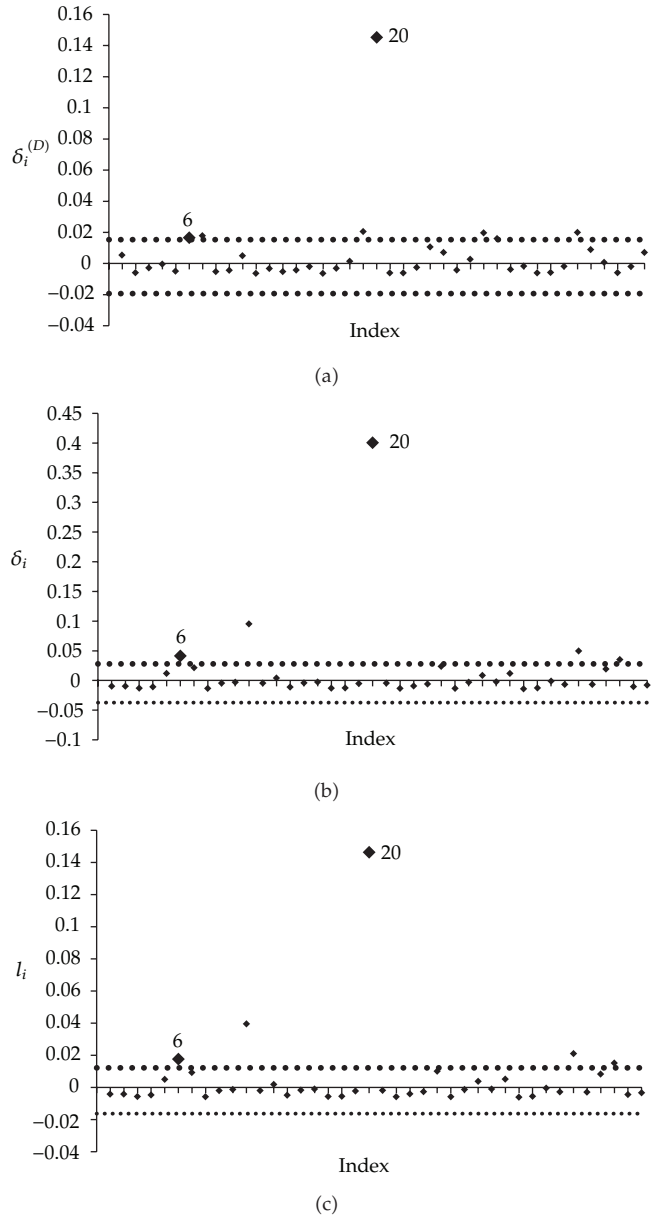


Figure 4: Index plot of collinearity-influential measures for original jet turbine engine data set.

300 with replication of 10,000 times were considered. Following the idea of Habshah et al. [7], two different contamination patterns were created. In the first pattern, 100 (α) percent observations of one of the generated collinear explanatory variables were replaced by high leverages with unequal weights. In this pattern the explanatory variable and the observation which needed to be replaced by high leverage point were chosen randomly. The second pattern is created by replacing the first 100 ($\alpha/2$) percent of one of the collinear explanatory variable and the last 100 ($\alpha/2$) percent of another collinear explanatory variable with high

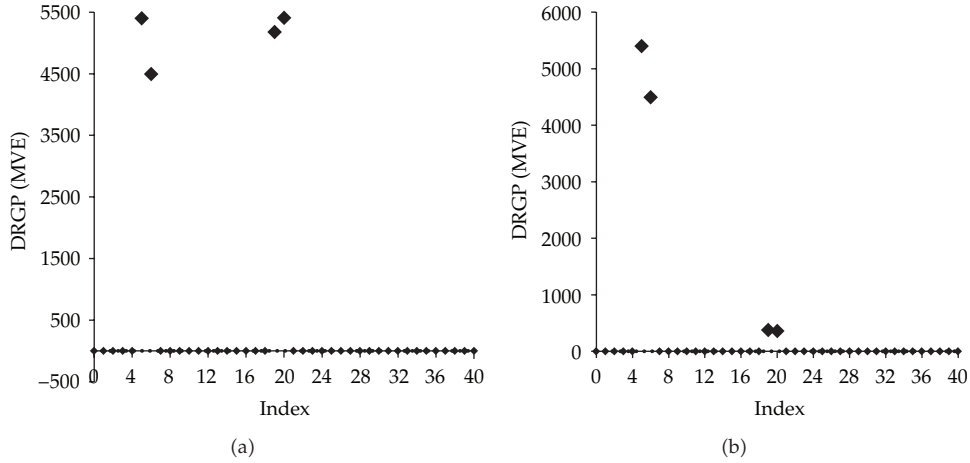


Figure 5: Index plot of DRGP (MVE) for modified jet turbine engine data set, (a) pattern1, (b) pattern2.

leverages with unequal weights. The two independent variables are also randomly selected and the replacement of the high leverage point to the observations in different positions of explanatory variables was also performed randomly. Following Habshah et al. [11] and Bagheri et al. [10], the high leverage values with unequal weights in these two patterns were generated such that the values corresponding to the first high leverage point are kept fixed at 10 and those of the successive values are created by multiplying the observations index, i by 10. The three diagnostic measures ${\delta_i}^{(D)}$, δ_i and l_i with the proposed cutoff point were introduced to (5.1) and were applied to each simulated data. The results based on the average values are presented in Table 2. The α and HLCIO in Table 2 indicate, respectively, the percentage and the number of added high leverage collinearity-reducing observations to the simulated data sets. Furthermore, the number of high leverage points which is detected by DRGP (MVE) is denoted as HL. It is interesting to point out that the percentage of the high leverage point, p^*_{ii} detected by DRGP (MVE) denoted as HL in Table 2 is more than the percentage of the added high leverage collinearity-reducing observations to the simulated data sets, α. However, by increasing the sample size and the percentage of added high leverage points to the simulated data, both percentages became exactly the same. The CN1 and the CN2 indicate the condition number of X matrix without and with high leverage collinearity-reducing observations, respectively. Moreover, $\text{Cut}(\theta_i)_1$ and $\text{Cut}(\theta_i)_2$ represent the number of high leverage collinearity-reducing observations and the number of collinearity-reducing observations which have been detected by cutoff θ_i.

Table 2 clearly shows the merit of our new proposed measure for high leverage collinearity-influential measure exhibited in (4.1). It can be observed that no other measures that were considered in this experiment performed satisfactorily except for our proposed measure. The simulated data sets have been created collinearly which produced large values of CN_1, condition number of simulated data sets without high leverage points ($\text{CN}_1 > 30$). The added multiple high leverage collinearity-influential observations reduces multicollinearity among the simulated explanatory variables; this reduction may result from the smaller values of CN_2 compared to CN_1. It is important mentioning that the reduction of the CN_2 values for the second pattern was much more significant compared to CN_2 for the first pattern. We can conclude that the influence of the added high leverage points to different

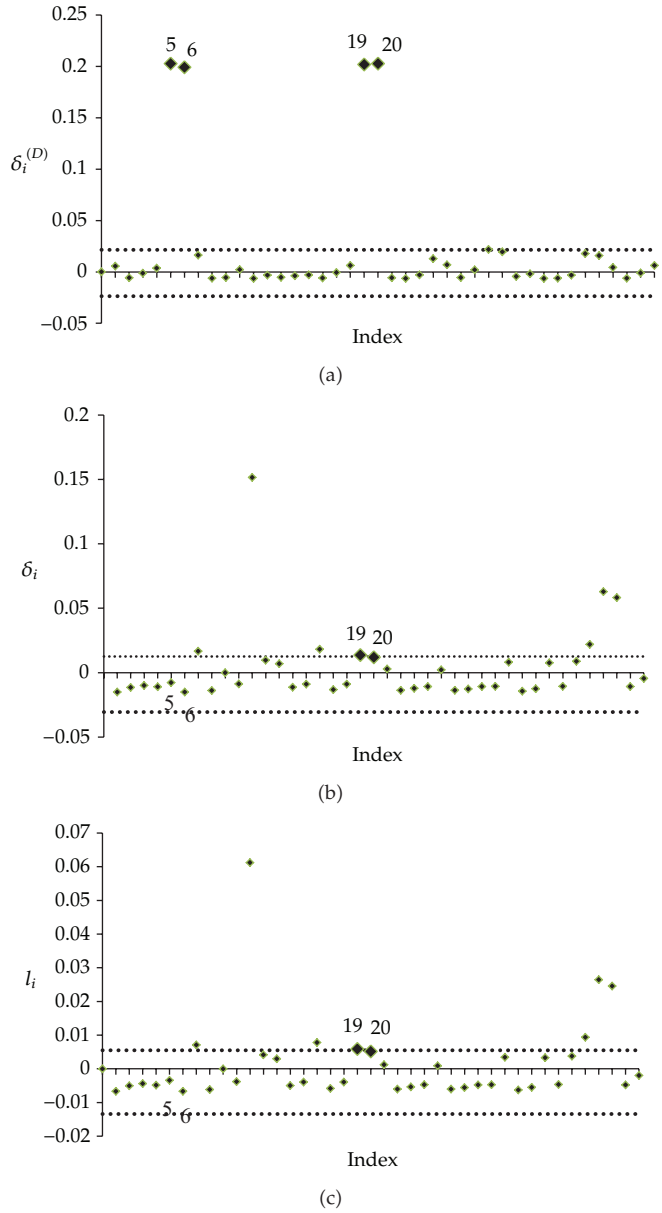


Figure 6: Index plot of collinearity-influential measures for the first modified pattern of jet turbine engine data set.

positions of two explanatory variables for changing the multicollinearity pattern of simulated data, is more significant compared to the added high leverage points to only one explanatory variable.

The results of Table 2 for the first pattern of simulated data sets indicate that for small sample sizes ($n = 20$) our proposed measure could not indicate the exact amount of high leverage collinearity-reducing observations. However, by increasing the sample size and the percentage of added high leverage points to the simulated data sets, the measure is capable

Table 2: Collinearity-influential measures for simulated data sets.

(a)

Measures	$n = 20$						$n = 40$					
	Pattern1			Pattern2			Pattern1			Pattern2		
α	10.00	20.00	30.00	10.00	20.00	30.00	10.00	20.00	30.00	10.00	20.00	30.00
HLCIO	2.00	4.00	6.00	2.00	4.00	6.00	4.00	8.00	12.00	4.00	8.00	12.00
HL	2.80	4.00	6.00	2.80	4.00	6.00	4.25	8.00	12.00	4.25	8.00	12.00
CN_1	49.53	53.27	51.72	49.53	53.27	51.72	40.91	40.33	40.02	40.91	40.33	40.02
CN_2	38.64	42.78	32.90	1.76	1.57	1.61	31.66	32.50	33.50	1.53	1.52	1.64
$\text{Cut}(\delta_i^{(D)})_1$	1.80	3.67	5.96	1.60	3.50	5.97	4.00	8.00	11.97	4.00	8.00	11.99
$\text{Cut}(\delta_i^{(D)})_2$	0.20	0.67	0.00	0.20	0.17	0.00	0.76	1.00	0.00	0.86	1.99	0.00
$\text{Cut}(\delta_i)_1$	0.00	0.17	0.99	1.20	0.00	0.00	0.04	0.00	0.01	0.00	0.00	0.00
$\text{Cut}(\delta_i)_2$	1.20	1.00	1.98	0.00	0.00	0.00	1.35	2.00	0.51	0.00	0.00	0.00
$\text{Cut}(l_i)_1$	0.00	0.17	0.99	1.20	0.00	0.00	0.04	0.00	0.01	0.00	0.00	0.00
$\text{Cut}(l_i)_2$	1.00	0.84	1.98	0.00	0.00	0.00	1.28	2.00	0.51	0.00	0.00	0.00

(b)

Measures	$n = 60$						$n = 100$					
	Pattern1			Pattern2			Pattern1			Pattern2		
α	10.00	20.00	30.00	10.00	20.00	30.00	10.00	20.00	30.00	10.00	20.00	30.00
HLCIO	6.00	12.00	18.00	6.00	12.00	18.00	10.00	20.00	30.00	10.00	20.00	30.00
HL	6.42	12.06	18.11	6.42	12.06	18.11	10.43	20.10	30.00	10.43	20.10	30.00
CN_1	38.58	37.62	39.72	38.58	37.62	39.72	36.87	38.08	37.97	36.87	38.08	37.97
CN_2	30.76	30.83	32.56	1.45	1.56	1.65	29.28	31.03	31.58	1.38	1.52	1.64
$\text{Cut}(\delta_i^{(D)})_1$	6.00	12.00	18.00	6.00	12.00	17.89	10.00	20.00	30.00	10.00	20.00	30.00
$\text{Cut}(\delta_i^{(D)})_2$	0.90	0.16	0.05	0.87	0.25	0.05	1.74	0.65	0.08	1.83	0.69	0.00
$\text{Cut}(\delta_i)_1$	0.19	0.40	0.58	0.00	0.00	0.00	0.34	0.42	0.99	0.00	0.00	0.00
$\text{Cut}(\delta_i)_2$	1.57	1.37	1.26	0.00	0.00	0.00	3.15	2.48	2.16	0.00	0.00	0.00
$\text{Cut}(l_i)_1$	0.19	0.40	0.58	0.00	0.00	0.00	0.26	0.38	0.99	0.00	0.00	0.00
$\text{Cut}(l_i)_2$	1.48	1.28	1.16	0.00	0.00	0.00	3.07	2.35	2.16	0.00	0.00	0.00

(c)

Measures	$n = 300$					
	Pattern1			Pattern2		
α	10.00	20.00	30.00	10.00	20.00	30.00
HLCIO	30.00	60.00	90.00	30.00	60.00	90.00
HL	30.00	60.00	90.00	30.00	60.00	90.00
CN_1	37.89	38.31	37.88	37.89	38.31	37.88
CN_2	30.92	31.60	32.23	1.34	1.49	1.66
$\text{Cut}(\delta_i^{(D)})_1$	30.00	60.00	90.00	30.00	60.00	90.00
$\text{Cut}(\delta_i^{(D)})_2$	3.61	0.80	0.00	4.09	0.73	0.00
$\text{Cut}(\delta_i)_1$	0.99	1.69	2.79	0.00	0.00	0.00
$\text{Cut}(\delta_i)_2$	6.63	6.58	7.13	0.00	0.00	0.00
$\text{Cut}(l_i)_1$	0.99	1.61	2.66	0.00	0.00	0.00
$\text{Cut}(l_i)_2$	6.63	6.58	7.13	0.00	0.00	0.00

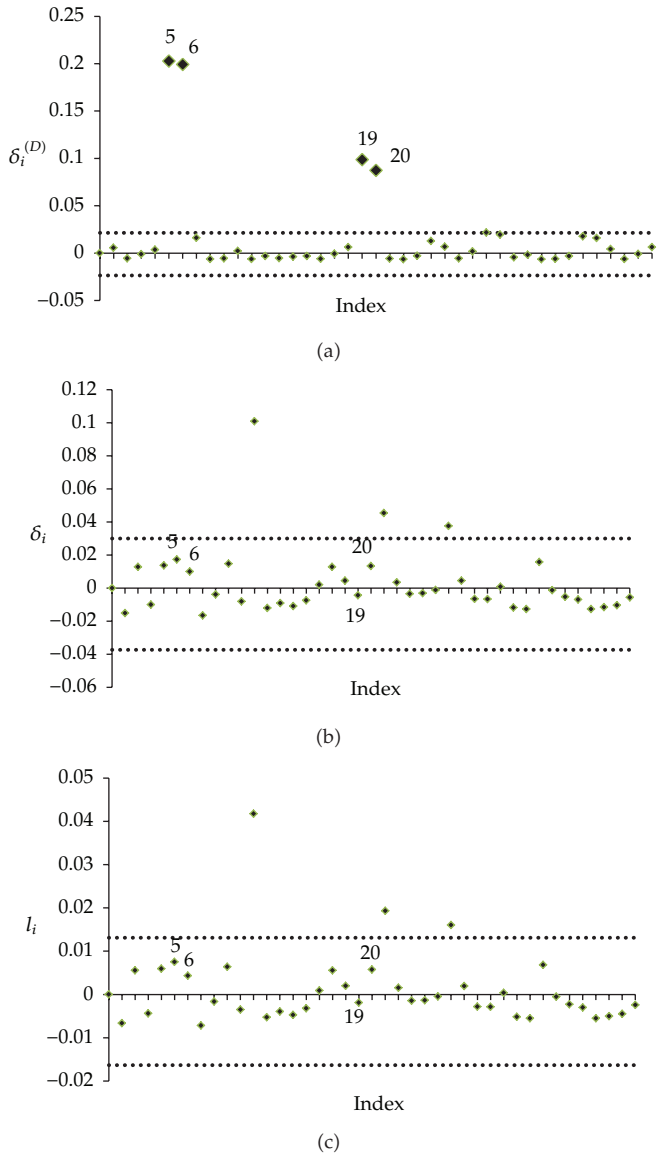


Figure 7: Index plot of collinearity-influential measures for the second modified pattern of jet turbine engine data set.

of detecting the exact amount of added high leverage collinearity-reducing observations. It is evident by looking at the value of $\text{Cut}(\delta_i^{(D)})_1$ is exactly the same as HLCIO. On the other hand, the other two collinearity-influential measures, δ_i and l_i, failed to indicate the exact amount of high leverage collinearity-reducing observations. It is worth noting that all of these three measures also detect some points as collinearity-reducing observations (see the $\text{Cut}(\theta_i)_2$ in Table 2, where θ_i is $\delta_i^{(D)}$, δ_i or l_i). Similar results will be obtained if pattern 1 can be drawn for the second pattern of the simulated data sets. Compared to the first contamination pattern, it is clearly seen that δ_i and l_i almost completely failed to detect either

high leverage collinearity-reducing observations or collinearity-reducing observations. Our proposed measure did a credible job where it is successfully detect high leverage collinearity-reducing observations for both contaminated patterns.

8. Conclusion

The presence of high leverage points and multicollinearity are inevitable in real data sets and they have an unduly effects on the parameter estimation of multiple linear regression models. These leverage points may be high leverage collinearity-enhancing or high leverage collinearity-reducing observations. It is crucial to detect these observations in order to reduce the destructive effects of multicollinearity on regression estimates which lead to misleading conclusion. It is easier to diagnose the presence of high leverage points which increase the collinearity among the explanatory variables compared to those which reduce collinearity. In this respect, it is very important to explore a sufficient measure with an accurate cutoff point for detecting high leverage collinearity-reducing observations. In this paper, we proposed a precise cutoff point for a novel existing measure to detect high leverage collinearity-reducing observations. By using an engineering data and a simulation study, we confirmed that the widely used measures failed to detect multiple high leverage collinearity-reducing observations. Furthermore, our proposed cutoff point successfully detects multiple high leverage collinearity-reducing observations.

References

[1] D. C. Montgomery, E. A. Peck, and G. G. Viving, *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York, NY, USA, 3rd edition, 2001.

[2] Md. Kamruzzaman and A. H. M. R. Imon, "High leverage point: another source of multicollinearity," *Pakistan Journal of Statistics*, vol. 18, no. 3, pp. 435–447, 2002.

[3] M. H. Kutner, C. J. Nachtsheim, and J. Neter, *Applied Linear Regression Models*, McGraw-Hill, New York, NY, USA, 2004.

[4] S. Chatterjee and A. S. Hadi, *Regression Analysis by Examples*, John Wiley & Sons, New York, NY, USA, 4th edition, 2006.

[5] A. S. Hadi, "Diagnosing collinearity-influential observations," *Computational Statistics & Data Analysis*, vol. 7, no. 2, pp. 143–159, 1989.

[6] M, A. Habshah, Bagheri, A. H. M. R, and Imon, "The application of robust multicollinearity diagnostic method based on robust coefficient determination to a non-collinear data," *Journal of Applied Sciences*, vol. 10, no. 8, pp. 611–619, 2010.

[7] M. Habshah, A. Bagheri, and A. H. M. R. Imon, "High leverage collinearity-enhancing observations and its effect on multicollinearity pattern; Monte Carlo simulation study," *Sains Malaysiana*, vol. 40, no. 12, pp. 1437–1447, 2011.

[8] A. Bagheri, H. Midi, and A. H. M. R. Imon, "The effect of collinearity-influential observations on collinear data set: A monte carlo simulation study," *Journal of Applied Sciences*, vol. 10, no. 18, pp. 2086–2093, 2010.

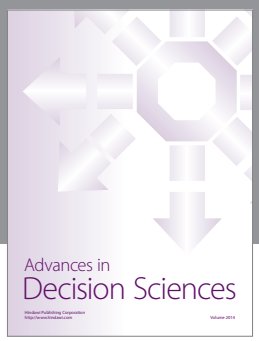
[9] A. Bagheri and H. Midi, "On the performance of robust variance inflation factors," *International Journal of Agricultural and Statistical Sciences*, vol. 7, no. 1, pp. 31–45, 2011.

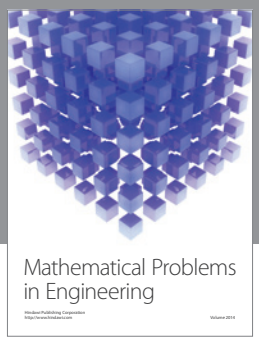
[10] A. Bagheri, M. Habshah, and R. H. M. R. Imon, "A novel collinearity-influential observation diagnostic measure based on a group deletion approach," *Communications in Statistics*, vol. 41, no. 8, pp. 1379–1396, 2012.

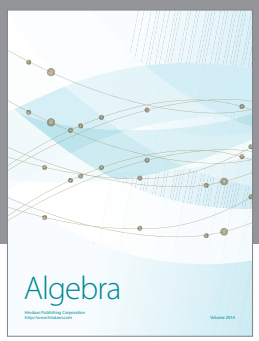
[11] M. Habshah, M. R. Norazan, and A. H. M. R. Imon, "The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression," *Journal of Applied Statistics*, vol. 36, no. 5-6, pp. 507–520, 2009.

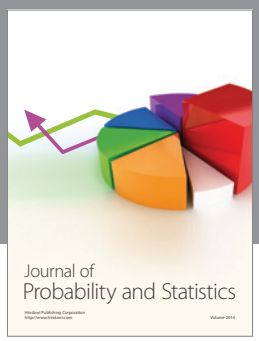
[12] D. Sengupta and P. Bhimasankaram, "On the roles of observations in collinearity in the linear model," *Journal of the American Statistical Association*, vol. 92, no. 439, pp. 1024–1032, 1997.

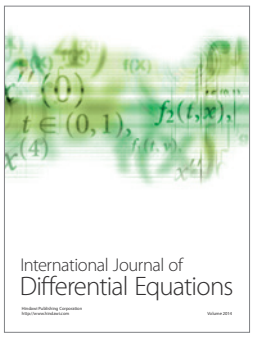
[13] A. S. Hadi, "A new measure of overall potential influence in linear regression," *Computational Statistics and Data Analysis*, vol. 14, no. 1, pp. 1–27, 1992.
[14] A. H. M. R. Imon, "Identifying multiple high leverage points in linear regression," *Journal of Statistical Studies*, vol. 3, pp. 207–218, 2002, Special Volume in Honour of Professor Mir Masoom Ali.
[15] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York, NY, USA, 1980.
[16] D. A. Belsley, *Conditioning Diagnostics-Collinearity and Weak Data in Regression*, John Wiley & Sons, New York, NY, USA, 1991.
[17] D. C. Hoaglin and R. E. Welsch, "The Hat Matrix in regression and ANOVA," *Journal of the American Statistical Association*, vol. 32, no. 1, pp. 17–22, 1978.
[18] P. Rousseeuw, "Multivariate estimation with high breakdown point," in *Mathematical Statistics and Applications*, pp. 283–297, Reidel, Dordrecht, The Netherlands, 1985.
[19] D. C. Montgomery, G. C. Runger, and N. F. Hubele, *Engineering Statistics*, John Wiley & Sons, New York, NY, USA, 5nd edition, 2011.
[20] G. W. Stewart, "Collinearity and least squares regression," *Statistical Science*, vol. 2, no. 1, pp. 68–100, 1987.
[21] K. D. Lawrence and J. L. Arthur, *Robust Regression; Analysis and Applications*, Marcel Dekker, New York, NY, USA, 1990.

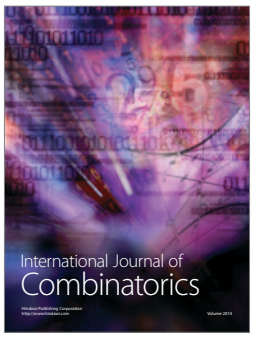


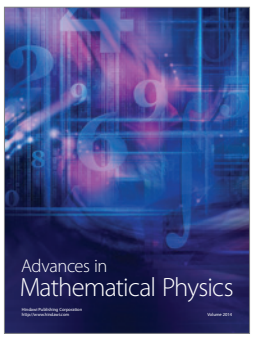


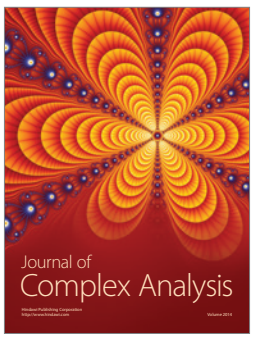


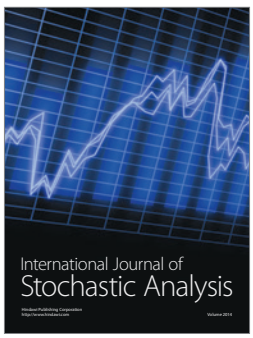


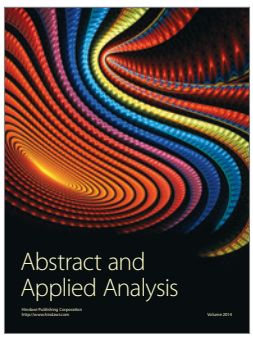


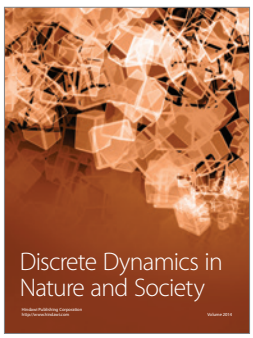


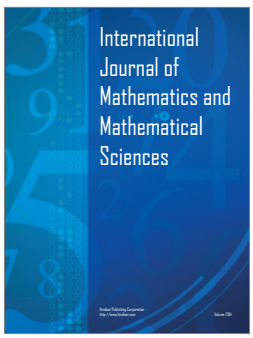


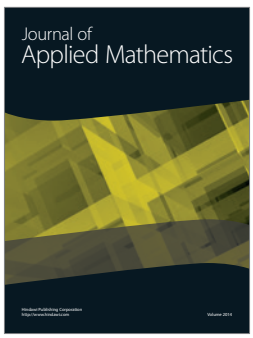












Hindawi
Submit your manuscripts at
http://www.hindawi.com