*International Journal of Statistical Sciences* *ISSN 1683–5603*
*Vol. 2, 2003, pp 37–49*
*© 2003 Dept. of Statistics, Univ. of Rajshahi, Bangladesh*

# A Solution to the Problem of Multicollinearity Caused by the Presence of Multiple High Leverage Points

**A. H. M. Rahmatullah Imon**
*Department of Statistics*,
University of Rajshahi , Bangladesh
Email: imon_ru@yahoo.com

**Md. Ashraful Islam Khan**
*Dept. of Population Science and Human Resource Development*,
University of Rajshahi, Bangladesh

[Received June 1, 2002; Revised October 2, 2002; Accepted June 11, 2003]

## Abstract

Multicollinearity often causes a huge interpretative problem in linear regression analysis. An important and almost inevitable, but not much focused source of multicollinearity is the existence of high leverage points. Omission of high leverage points from the analysis could be a remedy to this problem, but in the presence of multiple high leverage points, we anticipate that it may not be easy to eliminate the problem of multicollinearity. Generalised potentials have been in use to detect multiple high leverage points in linear regression. In this paper an attempt has been made to show how generalised potentials can be used as a remedy to multicollinearity problem. At first we present a few examples and figures, which draw our attention to this problem. Then we report a Monte Carlo simulation experiment designed to investigate how effective this technique is to solve the problem of multicollinearity caused by the presence of multiple high leverage points in linear regression.

**Keywords and Phrases:** Multicollinearity, high leverage points, masking, swamping, generalised potentials, Monte Carlo simulation.

## 1 Introduction

In a linear regression model, it is a convention to assume that there is no linear relationship among the regressors. Unfortunately in most applications, the regressors are nearly perfectly linearly related, and in such cases the inferences based on traditional methods become erroneous. When there are near linear dependencies among the regressors, the problem of multicollinearity is said to exit.

We write the multiple regression model as

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

where $Y$ is an $n \times 1$ vector of response or dependent variables, $X$ is an $n \times k$ $(n > k)$ matrix of predictors (explanatory variables) including one constant predictor, $\beta$ is a $k \times 1$ vector of unknown finite parameters to be estimated and $\in$ is an $n \times 1$ vector of random disturbances. Let the $j-$th column of the $X$ matrix be denoted $X_j$, so that $X = [X_1, X_2, \cdots, X_k]$. We define multicollinearity in terms of the linear dependence of the columns of $X$.

We generally use the ordinary least squares (OLS) technique to estimate the regression parameters $\beta$ because of tradition and ease of computation. But the presence of multicollinearity has a number of serious effects on the OLS estimates of the regression coefficients [see Montgomery and Peck (1992, pp. 291). A variety of sources of multicollinearity are now available in the literature [see Montgomery and Peck (1992, pp. 289). Kamruzzaman and Imon (2002) pointed out that the presence of high leverage points in a data set could be responsible for causing the problem of multicollinearity. In regression diagnostics we are more concerned with the identification of high leverage points together with outliers and influential observations. But in our work we deal with the cases where high leverage points are mainly responsible for causing multicollinearity. In section 2 we introduce the term leverage and briefly discuss some of the commonly used measures of leverages. However, it is now evident [see Imon (2002)] that in the presence of multiple high leverage points, many of them are masked in such a way that their identification becomes very difficult. We also introduce generalised potentials that were first proposed by Imon (1996) and then further studied [see Imon (2002)] for the identification of multiple high leverage points. We focus on the problem of multicollinearity caused by the presence of multiple high leverage points more extensively in section 3. We also show how the existing methods except generalised potentials fail to address this issue. We report a Monte Carlo simulation study in section 4, which is designed to investigate the role of generalised potentials as a remedy to the multicollinearity problem caused by the presence of multiple high leverage points.

## 2 High Leverage Points and Their Measures

In regression analysis it is sometimes very important to know whether any set of $X$-values are exerting too much influence on the fitting of the model. A set of influential $X$-values is known as a high leverage point. We can re-express the general linear model (1) by

$$y_i = x_i^T \beta + \in_i, \quad i = 1, 2, \cdots, n \tag{2}$$

where $y_i$ is the $i$-th observed response and $x_i$ is a $k \times 1$ vector of predictors. When the OLS method is employed to estimate the regression parameters we obtain $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$. Then the $i$-th residual is given by

$$\hat{\in}_i = y_i - x_i^T \hat{\beta}, \quad i = 1, 2, \cdots, n \tag{3}$$

In matrix notation this becomes

$$\hat{\in} = Y - X\hat{\beta} \tag{4}$$

which can also be expressed as

$$\hat{\in} = (I - W) \in \tag{5}$$

where $W = X \left(X^T X\right)^{-1} X^T$ which is generally known as weight matrix or leverage matrix. The diagonal elements of $W$, where

$$w_{ii} = x_i^T \left(X^T X\right)^{-1} x_i, \quad i = 1, 2, \cdots, n \tag{6}$$

are called the leverage values. Observations corresponding to excessively large $w_{ii}$ values are termed as high leverage points.

Much works have been done on the identification of high leverage points in linear regression. We know that the average value of $w_{ii}$ is $k/n$. Hoaglin and Welsch (1978) considered observations unusual when $w_{ii}$ exceeded $2k/n$ which is known as *twice-the-mean-rule*. Vellman and Welsch (1981) considered $w_{ii}$ as large when it exceeds $3k/n$ (*thrice-the-mean-rule*). For a definition of how large is a $w_{ii}$, Huber (1981, pp. 162) suggested breaking the range of possible values, $(0 \leq w_{ii} \leq 1)$ into three intervals. Values $w_{ii} \leq 0.2$ appear to be safe, values between 0.2 and 0.5 are risky, and values above 0.5 should be avoided. Well known Mahalanobis distances are also suggested to use as a measure of leverages in the literature, but Mahalanobis distance for each of the points has a one-to-one relationship with $w_{ii}$ [see Rousseeuw and Leroy (1987, pp. 224)].

Hadi (1992) pointed out that in the presence of a high leverage point the information matrix may break down and hence the observations may not have the appropriate leverages. He introduced a single case deleted measure of leverages known as potentials. We define the $i$-th potential as

$$p_{ii} = x_i^T \left(X_{(i)}^T X_{(i)}\right)^{-1} x_i \tag{7}$$

where $X_{(i)}$ is the data matrix $X$ with the $i$-th row deleted. However, it is easy to obtain a simple relationship between $w_{ii}$ and $p_{ii}$ as $p_{ii} = \frac{w_{ii}}{1-w_{ii}}$ [see Hadi (1992)]. Observations corresponding to excessively large potential values are considered as high leverage points. Hadi (1992) proposed a cut-off point for $p_{ii}$ as

$$\text{Mean}(p_{ii}) + c.\ \text{St. dev.}(p_{ii}) \tag{8}$$

where $c$ is an appropriately chosen constant such as 2 or 3. This form is analogous to a confidence bound for a location parameter. But the problem with this cut-off point is that both mean and variance of $p_{ii}$ may be non-robust in the presence of a single extreme value yielding a high cut-off point. To avoid such a problem Hadi (1992) suggested to replace the mean and the standard deviation in (8) by the median and the median absolute deviation (MAD) respectively.

It is reported by many authors [see Imon (2002)] that the presence of multiple high leverage points may cause masking and swamping and in that case a single case deletion method like potential is not enough to address this problem. Imon (1996) extends the idea of a single case deleted potential to a group deletion study. Let us denote a set of cases 'remaining' in the analysis by $R$ and a set of cases 'deleted' by $D$. Hence $R$ contains $(n-d)$ cases after $d < (n-k)$ cases in $D$ are deleted. Without loss of generality, assume that these observations are the last of $d$ rows of $X$ and $Y$. Using the result of Henderson and Searle (1981), we obtain

$$(X_R^T X_R)^{-1} = (X^T X)^{-1} + (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} X_D (X^T X)^{-1} \tag{9}$$

where $U_D = X_D (X^T X)^{-1} X_D^T$ is a symmetric matrix and $I_D$ is an identity matrix of order $d$. When a group of observations $D$ is omitted, we define weights for the entire data set as

$$w_{ii}^{(-D)} = x_i^T (X_R^T X_R)^{-1} x_i, \quad i = 1, 2, \cdots, n. \tag{10}$$

It should be noted that $w_{ii}^{(-D)}$ is the $i$-th diagonal element of $X(X_R^T X_R)^{-1} X^T$ matrix. Imon (1996) introduced generalised potentials for all members in a data set that are defined as

$$\begin{aligned} p_{ii}^* &= \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \quad \text{for} \quad i = 1, 2, \cdots, n-d \\ &= w_{ii}^{(-D)} \quad \text{for} \quad i = n-d+1, n-d+2, \cdots, n \end{aligned} \tag{11}$$

There exists no finite upper bound for $p_{ii}^\star$'s and it may not be easy to derive a theoretical distribution of them. But this does not make any problem to obtain a suitable confidence bound type cut-off point for them. One could consider $p_{ii}^\star$ to be large if

$$p_{ii}^* > \text{Median}(p_{ii}^*) + c\,\text{MAD}(p_{ii}^\star) \tag{12}$$

where $\text{MAD}(p_{ii}^*) = \text{Median}\{|p_{ii}^* - \text{Median}(p_{ii}^*)|\}/0.6745$.

Imon (2002) suggested a procedure for the identification of multiple high leverage points using generalised potentials to decide which points are of high leverages. As we observe from (11) that every value of $p_{ii}^*$ depends on the selection of the deletion set $D$, it is therefore important to be able to include all suspect cases into the $D$ set. For a $k$ variable regression, the $j$-th point of any regressor $X_i$ can be treated as suspect when it falls outside the interval

$$\text{Median}(X_i) \;\pm\; c\,\text{MAD}(X_i), \quad i = 1, 2, \cdots, k. \tag{13}$$

Not necessarily, the same data points (if any) of each regressor will satisfy rule (13). We would like to include all data points as members of the deletion set $D$ if they satisfy rule (13) for any $X_i$. But we have to impose a restriction on the maximum size of the deletion set $D$. The number of deleted observations by no means should exceed $n/2$, because if more than 50% observations are suspect then it is very difficult to distinguish high leverage points from low leverage points. At the same time we must make sure that the minimum number of observations remains in the analysis for the execution of the OLS technique. Thus the size of the remaining set $R$ should exceed $min(n/2, n-k)$. We then apply rule (12) to see whether all members of the deletion set have potentially high leverages or not. If all members of the $D$ set satisfy rule (12), we declare them as high leverage points. Since $p^*_{ii}$'s are measured in a similar scale, it should not matter too much if a low leverage point is included in the deletion set. But to be on the safe side, if members of the $D$ set do not satisfy rule (12), we prefer to put them back into the estimation subset $R$ sequentially (observation with the least $p^*_{ii}$ value will be replaced at the first) and to re-compute $p^*_{ii}$ values. We continue this process until all members of the deletion set individually satisfy rule (12). The points thus identified will be finally declared as high leverage points.

## 3 High Leverage Points and Muiticollinearity

It is now evident that high leverage points may cause multicollinearity in linear regression. If we are able to detect the high leverage points correctly we may get rid of the multicollinearity problem by deleting those observations. But we suspect that the commonly used detection techniques may fail to identify all of multiple high leverage points and the omission of observations thus identified may not help reduce the effect of multicollinearity. Here we present an example in favour of our proposition. We consider a well-known data set, which is frequently referred to the study of measuring influence of observations and identification of outliers. Hawkins *et al.* (1984) constructed this artificial data set containing 75 observations with 10 high leverage outliers (cases 1-10), 4 high leverage inliers (cases 11-14) and 61 low leverage inliers (cases 15-75). It is interesting to note that for this data set all of the commonly used measures of leverages fail to focus on most of the high leverage cases that in fact appear as points of low leverages.

Table 1 presents the commonly used leverage values $w_{ii}$ together with Hadi's potential values $p_{ii}$ and generalised potentials $p^*_{ii}$. It is clear from the results presented in this table that $w_{ii}$ values corresponding to the most of the high leverage points are not large enough and if any one considered 'twice-the-mean' rule only observations 12, 13 and 14 appear as the points of high leverages. Thrice-the-mean rule identifies only the 14th observation as high leverage point. Similar conclusion might be drawn following Huber (1981)'s suggestion. Though the $p_{ii}$ values are more sensitive to high leverage points this table shows that they fail to focus on the first 13 cases. When we apply rule (12) of the previous section we observe that the first 14 observations are appearing as points of high leverages. The generalised potential values presented in Table 1 are thus obtained from (11) with cases 1-14 deleted. This table also shows that the generalised potential values for the first 14 observations are clearly separated from the rest of the values.

Now we present various multicollinearity diagnostics for Hawkins *et al.* (1984) data. These results are presented in Table 2 that consider diagnostics for the original data set and the deleted data sets where high leverage points identified by twice-the-mean rule and generalised potentials are omitted. Several techniques have been proposed in the literature for detecting multicollinearity. Among them examination of the correlation matrix, variance inflation factor, tolerance, variance decomposition, examinations of eigen values, condition index and eigen value decomposition are very commonly used. For this particular data set we consider correlations, eigen values, variance inflation factors (VIF) and variance decompositions. For the original data we observe that the correlation coefficients between $X_1$, $X_2$ and $X_3$ are very high. We also observe two high variance inflation factors and two very low eigen values, which clearly indicate the presence of multicollinearity. Variance proportions corresponding to $X_2$ and $X_3$ also show that these two variables are affected by multicollinearity in the presence of $X_1$. As we suspect that the high leverage points are responsible for causing multicollinearity, their omission from the analysis should improve the situation. That is why we expect better results for the data set where deletion takes place on the leverage consideration. But we observe that the single case deleted diagnostic methods does not help identify high leverage points and consequently we observe a little improvement in the results of multicollinearity. But the use of generalised potentials produces stunning results. When the high leverage points identified by this method are omitted from the analysis, we observe that the values of correlation coefficients among $X_1$, $X_2$ and $X_3$ are very low. We also observe that neither of the VIF's is very high nor eigen values is very low. The results of variance proportions also show that there is no evidence of the presence of mulicollinearity in this data set and none of the three variables is affected by it.

Table 1: Leverages, potentials and generalised potentials for Hawkins *et al.* data

| Index | $w_{ii}$ | $p_{ii}$ | $p^*_{ii}$ | Index | $w_{ii}$ | $p_{ii}$ | $p^*_{ii}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.063 | 0.067 | $\underline{14.46}$ | 39 | 0.035 | 0.036 | 0.075 |
| 2 | 0.060 | 0.064 | $\underline{15.22}$ | 40 | 0.030 | 0.031 | 0.037 |
| 3 | 0.086 | 0.094 | $\underline{16.97}$ | 41 | 0.052 | 0.055 | 0.094 |
| 4 | 0.081 | 0.088 | $\underline{18.02}$ | 42 | 0.055 | 0.058 | 0.076 |
| 5 | 0.073 | 0.079 | $\underline{17.38}$ | 43 | 0.061 | 0.065 | 0.104 |
| 6 | 0.076 | 0.082 | $\underline{15.61}$ | 44 | 0.041 | 0.043 | 0.092 |
| 7 | 0.068 | 0.073 | $\underline{15.70}$ | 45 | 0.029 | 0.030 | 0.080 |
| 8 | 0.063 | 0.067 | $\underline{14.82}$ | 46 | 0.038 | 0.040 | 0.081 |
| 9 | 0.080 | 0.087 | $\underline{17.03}$ | 47 | 0.066 | 0.071 | 0.115 |
| 10 | 0.087 | 0.095 | $\underline{15.97}$ | 48 | 0.041 | 0.043 | 0.082 |
| 11 | 0.094 | 0.104 | $\underline{22.39}$ | 49 | 0.047 | 0.049 | 0.062 |
| 12 | $\underline{0.144}$ | $\underline{0.168}$ | $\underline{24.03}$ | 50 | 0.016 | 0.016 | 0.056 |
| 13 | $\underline{0.109}$ | $\underline{0.122}$ | $\underline{22.73}$ | 51 | 0.036 | 0.037 | 0.058 |
| 14 | $\underline{0.564}$ | $\underline{1.294}$ | $\underline{28.16}$ | 52 | 0.072 | 0.078 | 0.098 |
| 15 | 0.058 | 0.062 | 0.091 | 53 | 0.079 | 0.086 | 0.139 |
| 16 | 0.076 | 0.082 | 0.104 | 54 | 0.040 | 0.042 | 0.083 |
| 17 | 0.039 | 0.041 | 0.086 | 55 | 0.034 | 0.035 | 0.051 |
| 18 | 0.023 | 0.024 | 0.027 | 56 | 0.037 | 0.039 | 0.066 |
| 19 | 0.031 | 0.032 | 0.046 | 57 | 0.023 | 0.024 | 0.050 |
| 20 | 0.048 | 0.050 | 0.096 | 58 | 0.040 | 0.042 | 0.072 |
| 21 | 0.029 | 0.030 | 0.036 | 59 | 0.019 | 0.019 | 0.046 |
| 22 | 0.046 | 0.048 | 0.072 | 60 | 0.062 | 0.066 | 0.099 |
| 23 | 0.029 | 0.030 | 0.041 | 61 | 0.051 | 0.054 | 0.111 |
| 24 | 0.026 | 0.027 | 0.047 | 62 | 0.021 | 0.021 | 0.090 |
| 25 | 0.022 | 0.022 | 0.089 | 63 | 0.036 | 0.037 | 0.076 |
| 26 | 0.032 | 0.033 | 0.069 | 64 | 0.026 | 0.027 | 0.080 |
| 27 | 0.042 | 0.044 | 0.090 | 65 | 0.031 | 0.032 | 0.060 |
| 28 | 0.024 | 0.025 | 0.035 | 66 | 0.036 | 0.037 | 0.055 |
| 29 | 0.018 | 0.018 | 0.039 | 67 | 0.019 | 0.019 | 0.022 |
| 30 | 0.047 | 0.049 | 0.100 | 68 | 0.046 | 0.048 | 0.099 |
| 31 | 0.059 | 0.057 | 0.070 | 69 | 0.029 | 0.030 | 0.072 |
| 32 | 0.036 | 0.037 | 0.073 | 70 | 0.027 | 0.028 | 0.050 |
| 33 | 0.026 | 0.027 | 0.046 | 71 | 0.019 | 0.019 | 0.034 |
| 34 | 0.032 | 0.033 | 0.094 | 72 | 0.028 | 0.029 | 0.032 |
| 35 | 0.034 | 0.035 | 0.082 | 73 | 0.043 | 0.045 | 0.048 |
| 36 | 0.023 | 0.024 | 0.040 | 74 | 0.050 | 0.053 | 0.058 |
| 37 | 0.059 | 0.063 | 0.092 | 75 | 0.062 | 0.066 | 0.096 |
| 38 | 0.021 | 0.021 | 0.056 | | | | |

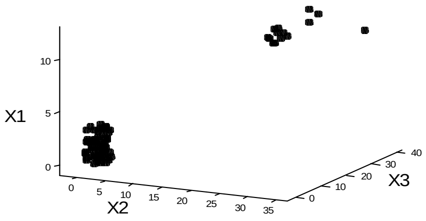

Figure 1. 3D plot of the original *X*'s of Hawkins *et al*. data

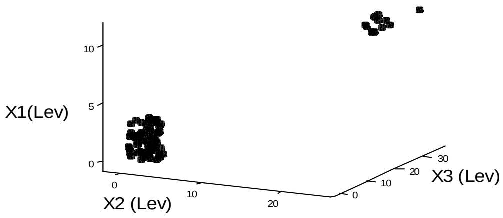

Figure 2. 3D plot of the *X*'s after deleting the cases by 2M method for Hawkins *et al*. data

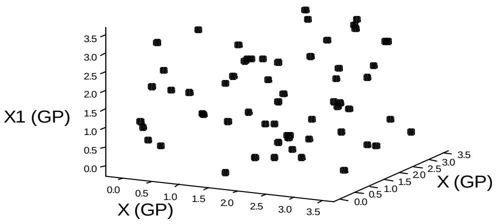

Figure 3. 3D plot of the *X*'s after deleting the cases by GP method for Hawkins *et al*. data

Table 2: Multicollinearity diagnostics for Hawkins *et al.* data

| Data | Correlation | Eigen Value | VIF | Variance Proportion | | |
|---|---|---|---|---|---|---|
| | | | | $X_1$ | $X_2$ | $X_3$ |
| Original ($n = 75$) | $r_{12} = 0.946$ | 3.369 | 2.402 | 0.00 | 0.00 | 0.00 |
| | $r_{13} = 0.962$ | 0.584 | 10.026 | 0.80 | 0.28 | 0.02 |
| | $r_{23} = 0.979$ | 0.034 | 15.997 | 0.20 | 0.72 | 0.98 |
| | | 0.013 | | | | |
| Del. Lev. ($n = 72$) | $r_{12} = 0.945$ | 3.352 | 2.364 | 0.00 | 0.00 | 0.00 |
| | $r_{13} = 0.951$ | 0.600 | 9.320 | 0.97 | 0.08 | 0.05 |
| | $r_{23} = 0.987$ | 0.039 | 19.091 | 0.03 | 0.92 | 0.95 |
| | | 0.009 | | | | |
| Del. GP. ($n = 61$) | $r_{12} = 0.044$ | 3.383 | 3.434 | 0.76 | 0.22 | 0.07 |
| | $r_{13} = 0.107$ | 0.287 | 3.823 | 0.03 | 0.44 | 0.68 |
| | $r_{23} = 0.127$ | 0.232 | 5.862 | 0.21 | 0.34 | 0.25 |
| | | 0.098 | | | | |

Now we present some 3D plots of explanatory variables that will show how generalised potentials contribute in handling the multicollinearity problem. Figure 1 presents a 3D plot of the $X$'s with the original data. As we know that 14 out of 75 observations are high leverage points we observe a strong indication of the presence of multicollinearity in the data. We observe similar picture in Figure 2 where 3 out of 14 high leverage points are omitted. Figure 3 presents a 3D plot of the $X$'s where high leverage points detected by generalised potential method are omitted. This plot clearly shows no sign of multicollinearity that reemphasises our view that the problem of multicollinearity could be eliminated if all of the genuine high leverage points are omitted from the analysis.

## 4 Simulation Results

In this section we report a Monte Carlo simulation study, which is designed to investigate how high leverage points behave as a source of multicollinearity. We consider an artificial two-predictor data set where our $X$'s are generated independently as Uniform (0,1) so that the correlation coefficient becomes very low (near to 0). We then deliberately change or set some values of $X'$s so that they become points of high leverages. At first we consider cases where high leverage points have equal weight. We also investigate the performances of different methods for detecting high leverage points in the process of multicollinearity reduction when high leverage points thus identified are omitted from the regression model. Then we extend this experiment to the cases where multiple high leverage points have different weights. We generate artificial data sets for multiple (10%) high leverage cases with equal and unequal weights. For both of the designs we consider cases for six different sample sizes ($n$ = 20, 30, 40, 50, 100 and 200) and six high leverage points ($x$ = 2, 3, 4, 5, 8, 10) and then the correlation coefficient for these two regressors are computed. The correlation method is certainly not the best way of detecting multicollinearity but we use this method because it is very simple, easy to compute and higher correlation always guarantee multicollinearity [Belsley (1991, pp. 20)]. Throughout our simulation experiment we use five detection techniques, twice the mean rule (2M), thrice the mean rule (3M), Huber method [cut-off rule ($w_{ii} > 0.2$)], Hadi's potential [cut-off rule $p_{ii} >$ Median($p_{ii}$) + 3MAD($p_{ii}$)] and Imon's generalised potential methods to identify high leverage points. Correlation coefficients are computed after omitting the observations identified by these five detection techniques. The results of these experiments are presented in Tables 3 and 4, each of which is based on 10000 simulations.

First we present the simulation results where the $X$ variables contain 10% equal high leverage points. For each of the cases the first 90% observations are simulated as Uniform (0,1). The last 10% observations of $X_1$and $X_2$ are set at six set of high $x$ values (i.e. $x$ = 2, 3, 4, 5, 8 and 10) so that these points are considered as high leverage points with equal weights. Correlation coefficients of $X'$s together with the results after excluding the suspect high leverage cases by different detection techniques are presented in Table 3.

We observe from Table 3 that for every $n$, the presence of multiple high leverage points causes strong multicollinearity. It is interesting to note that the correlations between the $X$'s tend to reduce slightly with the increase in sample size but the correlation tends to increase with the increase in leverage values. Throughout the simulations we observe that the performance of 3M is very poor. We observe no improvement of using this technique in multicollinearity reduction. Huber's method is appeared to be good for small samples, but even for the moderate sample size like 50 it breaks down. Potential method is also good for small samples, but its performance tends to deteriorate with the increase in sample size. The performance of 2M rule is satisfactory for all samples but throughout the simulation experiment generalised potentials over perform the rest of the methods considered in this study.

Next we report another simulation experiment where the $X$ variables contain 10% high leverage points having unequal weights. For each of the cases the first 90% observations are simulated as Uniform (0,1). The last 10% observations of $X_1$ and $X_2$ are taken serially from a set of observations starting from 2 and then having increments of 2 (i.e. $x$ = 2, 4, 6, 8, 10,$\cdots$, 40) so that these points are considered as high leverage points with unequal weights. Correlation coefficients of $X'$s together with the results after excluding the suspect high leverage cases by different detection techniques are computed and these results are presented in Table 4.

We observe from results of Table 4 that the presence of multiple unequal high leverage points causes strong multicollinearity, even stronger than the equal high leverage cases. Likewise the previous experiment the correlations between the $X$'s tend to in-

Table 3: Correlation coefficients of the $X$'s with 10% equal high leverage points

| Sample size | Measures | Correlation | | | | | |
|---|---|---|---|---|---|---|---|
| | | $x$ =2 | $x$ =3 | $x$ =4 | $x$ =5 | $x$ =8 | $x$ =10 |
| $n$= 20 | ACTUAL | 0.7484 | 0.8890 | 0.9411 | 0.9630 | 0.9861 | 0.9915 |
| | 2M | 0.0358 | 0.0203 | 0.0351 | 0.0300 | 0.0192 | 0.0333 |
| | 3M | 0.7487 | 0.8890 | 0.9411 | 0.9630 | 0.9861 | 0.9915 |
| | Huber | 0.0292 | 0.2789 | 0.2928 | 0.2945 | 0.2761 | 0.2906 |
| | Potential | 0.0862 | 0.0377 | 0.0503 | 0.0450 | 0.0288 | 0.0112 |
| | GP | 0.0160 | -0.0051 | 0.0180 | 0.0090 | -0.0006 | 0.0032 |
| $n$= 30 | ACTUAL | 0.7369 | 0.8868 | 0.9390 | 0.9621 | 0.9860 | 0.9913 |
| | 2M | 0.0328 | 0.0179 | 0.0292 | 0.0201 | 0.0184 | 0.0181 |
| | 3M | 0.7394 | 0.8869 | 0.9389 | 0.9621 | 0.9859 | 0.9912 |
| | Huber | 0.0951 | 0.0788 | 0.0935 | 0.0912 | 0.0859 | 0.0845 |
| | Potential | 0.1822 | 0.0892 | 0.0740 | 0.0641 | 0.0526 | 0.0497 |
| | GP | 0.0086 | 0.0018 | 0.0072 | 0.0025 | -0.0026 | 0.0030 |
| $n$= 40 | ACTUAL | 0.7357 | 0.8862 | 0.9380 | 0.9617 | 0.9858 | 0.9912 |
| | 2M | 0.0219 | 0.0240 | 0.0133 | 0.0138 | 0.0212 | 0.0204 |
| | 3M | 0.7372 | 0.8863 | 0.9380 | 0.9617 | 0.9858 | 0.9911 |
| | Huber | 0.0219 | 0.0240 | 0.0133 | 0.0138 | 0.0212 | 0.0204 |
| | Potential | 0.2837 | 0.1504 | 0.0970 | 0.1164 | 0.1064 | 0.1084 |
| | GP | 0.0025 | 0.0053 | -0.0047 | -0.0017 | 0.0030 | 0.0050 |
| $n$= 50 | ACTUAL | 0.7347 | 0.8846 | 0.9378 | 0.9614 | 0.9859 | 0.9911 |
| | 2M | 0.0119 | 0.0152 | 0.0153 | 0.0169 | 0.0233 | 0.0208 |
| | 3M | 0.7348 | 0.8846 | 0.9378 | 0.9614 | 0.9859 | 0.9911 |
| | Huber | 0.7353 | 0.8850 | 0.9381 | 0.9615 | 0.9859 | 0.9911 |
| | Potential | 0.3489 | 0.2085 | 0.1894 | 0.1851 | 0.1745 | 0.1612 |
| | GP | -0.0018 | -0.0008 | 0.0005 | 0.0012 | 0.0104 | 0.0046 |
| $n$= 100 | ACTUAL | 0.7335 | 0.8839 | 0.9374 | 0.9610 | 0.9857 | 0.9909 |
| | 2M | 0.0212 | 0.0122 | 0.0166 | 0.0122 | 0.0193 | 0.0242 |
| | 3M | 0.7335 | 0.8839 | 0.9374 | 0.9610 | 0.9857 | 0.9909 |
| | Huber | 0.7335 | 0.8839 | 0.8970 | 0.9610 | 0.9857 | 0.9909 |
| | Potential | 0.5681 | 0.4428 | 0.4065 | 0.3849 | 0.3801 | 0.3689 |
| | GP | 0.0067 | 0.0001 | 0.0056 | -0.0013 | 0.0059 | 0.0010 |
| $n$= 200 | ACTUAL | 0.7308 | 0.8830 | 0.9366 | 0.9606 | 0.9856 | 0.9909 |
| | 2M | 0.0106 | 0.0101 | 0.0108 | 0.0112 | 0.0152 | 0.0149 |
| | 3M | 0.7308 | 0.8830 | 0.9366 | 0.9606 | 0.9856 | 0.9909 |
| | Huber | 0.7308 | 0.8830 | 0.9366 | 0.9606 | 0.9856 | 0.9909 |
| | Potential | 0.6864 | 0.6675 | 0.6197 | 0.6187 | 0.5810 | 0.5773 |
| | GP | 0.0006 | 0.0012 | 0.0008 | -0.0013 | 0.0041 | 0.0033 |

Table 4: Correlation coefficients of the $X$'s with 10% unequal high leverage points

| Measures | Correlation | | | | | |
|---|---|---|---|---|---|---|
| | $n$ =20 | $n$ = 30 | $n$ = 40 | $n$ = 50 | $n$ = 100 | $n$ = 200 |
| ACTUAL | 0.9042 | 0.9498 | 0.9698 | 0.9799 | 0.9945 | 0.9986 |
| 2M | 0.5178 | 0.6204 | 0.8345 | 0.9252 | 0.9626 | 0.9896 |
| 3M | 0.6066 | 0.8631 | 0.9356 | 0.9217 | 0.9766 | 0.9950 |
| Huber | 0.5811 | 0.8120 | 0.8345 | 0.9221 | 0.9896 | 0.9985 |
| Potential | 0.6370 | 0.6755 | 0.8342 | 0.9094 | 0.9674 | 0.9920 |
| GP | 0.0056 | 0.0041 | 0.0068 | -0.0029 | 0.0016 | 0.0015 |

crease with the increase in leverage values. For this experiment both the number and magnitude of leverage values go up with the increase in sample size which also lead to higher correlation. But it is interesting to note that all detection techniques except the generalised potentials break down completely in the presence of unequal high leverage points. So far successful 2M method also breaks down here. Even for a small sample size like $n = 20$, we observe little improvement of using this technique in the multicollinearity reduction and its performance tends to deteriorate with the increase in sample size. Similar remarks may go with 3M, Huber and potential methods. But the performance of generalised potentials is quite outstanding. For all samples we observe that the omission of the cases identified by this method produces very low correlation coefficients.

## 5 Conclusions

The handling of multicollinearity generated by the presence of high leverage points in a linear regression model is investigated in this paper. One simple remedy is to detect the high leverage points and then to fit the model excluding them. But from the examples, figures and simulation results we observe that the detection and consequently the multicollinearity reduction process may be extremely complicated in the presence of multiple high leverage points. For multiple high leverage points with equal weights we observe that the performance of generalised potentials is the best followed by twice-the-mean rule. The performances of Huber and potential methods are good only for small samples. Thrice-the-mean rule performs very poorly in this case. But in the presence of multiple high leverage points with unequal weights masking/swamping may occur. That is, most of the commonly used detection methods may fail to identify all of the high leverage points. In that case, multicollinearity is reduced though the problem still remains unresolved. On the contrary performance of generalised potentials is quite outstanding. Irrespective of sample size and leverage structure its performance is very robust. We observe that the omission of the cases identified by this method can remove the multicollinearity effect from the data.

## References

Belsley, D.A. (1991). *Conditioning Diagnostics; Collinearity and Weak Data in Regression*, Wiley, New York.

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

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

Henderson, H.V. and Searle, S.R. (1981). On deriving the inverse of a sum of matrices, *SIAM Review*, 22, 53-60.

Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA, *Journal of the American Statistical Association* 32, 17-22.

Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

Imon, A.H.M.R. (1996). *Subsample Methods in Regression Residual Prediction and Diagnostics*, Ph.D. thesis, School of Mathematics and Statistics, University of Birmingham, U.K.

Imon, A.H.M.R. (2002). Identifying multiple high leverage points in linear regression, *Journal of Statistical Studies*, Special Volume in Honour of Professor Mir Masoom Ali, 207-218.

Kamruzzaman, Md. and Imon, A.H.M.R. (2002). High leverage point: Another source of multicollinearity, *Pakistan Journal of Statistics*, 18, 435-448.

Montgomery, D.C. and Peck, E,A. (1992). *Introduction to Linear Regression Analysis*, 2nd Ed., Wiley, New York.

Rousseeuw, P.J. and Leroy, A. (1987). *Robust Regression and Outlier Detection,* Wiley, New York.

Velleman, P.F. and Welsch, R.E. (1981). Efficient computing of regression diagnostics, *The American Statisticians*, 35, 234-242.