2003

The Role of High Leverage Points in Regression Diagnostics

Khan, Md. Ashraful Islam

University of Rajshahi

THE ROLE OF HIGH LEVERAGE POINTS IN REGRESSION DIAGNOSTICS

A
Dissertation
Submitted to the University of Rajshahi in
Fulfillment of the Requirements for the degree of
Master of philosophy

By

Md. Ashraful Islam Khan

NOVEMBER, 2003

DEPARTMENT OF STATISTICS
UNIVERSITY OF RAJSHAHI
RAJSHAHI, BANGLADESH.

Certificate

I am pleased to certify that Md Ashraful Islam Khan, Lecturer, Department of Population Science & Human Resource Development, University of Rajshahi for submission of the M. Phil. Thesis entitled "The Role of High Leverage Points in Regression Diagnostics".

I do hereby certify that the works embodied in this dissertation were carried out by the candidate. His work is original and genuine. No part of this study has been submitted in substance for any higher degree or diploma.

I wish his success.

A. H. M. Rahmatullah Imon
19.11.03

(Dr. A H. M. Rahmatullah Imon)

Supervisor

Associate Professor

Department of Statistics

University of Rajshahi,

Bangladesh

Statement of Originality

This dissertation does not incorporate any part without acknowledgement of any material previously submitted for a higher degree or diploma in any University or Institution and to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

19.11.03

(Md. Ashraful Islam Khan)

Lecturer,
Dept. of Population Science & Human Resource Development
University of Rajshahi
Bangladesh.

University of Rajshahi
November, 2003

DEDICATED
TO
MY PARENTS
AND
FRIEND

Acknowledgement

I would like to thank my supervisor ***Dr. A.H.M. Rahmatullah Imon*** *of the Department of Statistics, University of Rajshahi for his constant inspiration, constructive guidance and help throughout my research work for which I am in a position to submit this thesis*

I gratefully acknowledge all of my honourable teachers and colleagues.

Finally, I would like to thank my family members especially to my parents for their encouragement and supports. I would also like to thank all of my beloved friends.

Md. Ashraful Islam Khan

Synopsis

In fitting a linear regression model by the least squares method, leverage values play a very important role. They often form the basis of regression diagnostics as measures of influential observations in the explanatory variables. Much work have been done on the detection of high leverage values and a good number of diagnostic measures are now available in the literature. But neither of these methods is effective in the identification of high leverage points when multiple high leverage points are present in the data. In our study we proposed a new method for the identification of multiple high leverage points. The usefulness of this newly proposed method is studied under a variety of leverage structures through Monte Carlo simulation experiments. We also investigated the performance of the newly proposed method as a remedy to multicollinearity problem caused by the presence of multiple high leverage points.

List of Contents

No.	Title	Page No.
Chapter One: Introduction	.....	1-4
Chapter Two: Diagnostics in Linear Regression	.....	5-27
2.1	Regression Analysis	5
2.2	Historical Origin of the Term "Regression"	7
2.3	The Most Popular Regression Technique	7
2.4	Principles of Ordinary Least Squares Method	8

2.4.1 The Ordinary Least Square Estimator of the Regression Coefficient β 10

2.5 The Ordinary Least Squares (OLS) Residuals 11

2.6 The Weight Matrix, W 14

2.7 Regression Diagnostics 14

2.7.1 Departure from Classical Assumptions 14

2.7.2 The Normality Assumption 15

2.7.3 Why Testing for Normality Assumption 16

2.7.4 Problems for Departure from the Normality Assumption 16

2.8 Influential Observations, High Leverage Points and Outliers . 18

2.8.1 Outliers 18

2.8.2 Influential Observations 20

2.8.3 High Leverage Points 21

2.8.4 Inter Relationships Among Outliers, Influential Observations and High Leverage Points 22

2.8.5 Consequence of the Presence of Outliers, High Leverage Points and Influential Observations 25

2.9 Robust Regression Techniques 26

Chapter Three: Measures of Leverages 28-45

3.1 Masking and Swamping 28

3.2 Measures of Leverages 31

3.2.1 Properties of Weight Matrix, W 31

3.2.2 Different Measures of Leverages 37

3.2.3 Relation Between Mahalanobis Distance and Leverage Values..... 42

3.3 Comparison Between Potentials and Leverage Values..... 44

Chapter Four: Identification of a Single High Leverage Point 46-67

4.1 Simulation 47

4.2 Sensitivity of Different Measures of Leverages 51

4.2.1 Result Discussion for no High Leverage Cases.. 54

4.2.2 Simulation Results for Different Sample Sizes .. 54

4.2.3 Simulation Result Discussion for no High Leverage Cases 56

4.3 Identification of a Single High Point 57

4.3.1 Result Discussion for Single High Leverage Case 60

4.3.2 Simulation Results for Different Measures of Leverages..... 60

4.3.3 Simulation Results Discussion for Single High Leverage Case..... 65

Chapter Five: Identification of Multiple High Leverage Points 68-95

5.1 Generalized Potential 69

5.2 Identification of Multiple (10%) Equally High Leverage Points 77

5.2.1 Different Examples for the Performance of the Seven Sets of Measures 77

5.2.2 Result Discussion 81

5.2.3 Simulation Results 81

5.2.4 Simulation Results Discussion 91

5.3 Identification of Multiple (10%) Unequally High Leverage Points 86

5.3.1 Simulation Result Discussion 87

5.4 Graphical Display for Locating Multiple High Leverage Points 87

5.4.1 Leverage-Residual (L-R) plot 88

5.4.2 Potential-Residual (P-R) plot 88

5.4.3 Generalized Potential-Dilation Residual (GP-DR) Plot..... 89

Chapter Six: Multicillinearity and High Leverage Points 96-122

6.1 Concept of Multicollinearity 97

6.2 The Nature of Multicollinearity 97

6.3 Source of Multicollinearity 99

6.4 Effects or Consequences of Multicolliearity 100

6.5 Detection Techniques of Multicollinearity 100

6.6 Methods of Dealing with multicollinearity 104

6.7 High Leverage Points and Multicollinearity 109

6.8 Simulation Results 112

6.81 Simulation Results for a Single High Leverage Point 113

6.8.2 Simulation Results Discussion for a Single High Leverage Point 117

6.8.3 Simulation Results for 10% equal High Leverage Point 117

6.8.4 Simulation Results Discussion for 10% equal High Leverage Point 121

6.8.5 Simulation Results for 10% High Leverage Point 121

6.8.6 Simulation Results Discussion for 10% High Leverage Point 122

Chapter Seven: Conclusion and Areas of Further Research 123-125

7.1 Discussion of Results 123

7.2 Areas of Further Research 125

Appendix 126-137

References 138-144

List of Tables

No	Title	Page No.
3.1	Modified Peña and Yohai (1995) Data.....	44
4.a.1	Results for Different Measure of Leverages for n=10	51
4.a.2	Results for Different Measure of Leverages for n=20	52
4.a.3	Results for Different Measure of Leverages for n=40	53
4.b	Swamping of Cases Using Different Measures of Leverages	55
4.c.1	Results of Different Measure of Leverages in Presence of Single Leverage Point for n=10	57
4.c.2	Results of Different Measure of Leverages in Presence of Single Leverage Point for n=20	58

4.c.3 Results of Different Measure of Leverages in Presence of Single Leverage Point for n=40 59

4.d.1 Simulation Result on the Identification of a Single High Leverage Point for n=10 61

4.d.2 Simulation Result on the Identification of a Single High Leverage Point for n=20 62

4.d.3 Simulation Result on the Identification of a Single High Leverage Point for n=30 62

4.d.4 Simulation Result on the Identification of a Single High Leverage Point for n=40 63

4.d.5 Simulation Result on the Identification of a Single High Leverage Point for n=50 63

4.d.6 Simulation Result on the Identification of a Single High Leverage Point for n=100 64

4.d.7 Simulation Result on the Identification of a Single High Leverage Point for n=200 64

5.a.1 Hawkins-Bradu-Kass (1984) Artificial Data 72

5.a.2 Leverages and Potentials and Generalized Potential for Hawkings-Bradu-Kass (1984) Artificial Data 73

5.b.1 Results of Measure of Leverages 10% High Leverage Point for n=20 . 78

5.b.2 Results of Measure of Leverages 10% High Leverage Point for n=30 ... 79

5.b.2 Results of Measure of Leverages 10% High Leverage Point for n=40 . 80

5.c.1 Simulation Results in Presence of Multiple (10%) Equally High Leverage Point (2) 82

5.c.2 Simulation Results in Presence of Multiple (10%) Equally High Leverage Point (3) 82

5.c.3 Simulation Results in Presence of Multiple (10%) Equally High Leverage Point (4) 83

5.c.4 Simulation Results in Presence of Multiple (10%) Equally High Leverage Point (5) 83

5.c.5 Simulation Results in Presence of Multiple (10%) Equally High Leverage Point (8) 84

5.c.6 Simulation Results in Presence of Multiple (10%) Equally High Leverage Point (10) 84

5.d Simulation Results in Presence of Multiple (10%) Unequally High Leverage Points 86

5.e OLS and RLS Residuals for Hawkins-Bradu-Kass (1984) Data 91

5.f Leverages and Residuals for Peña and Yohai(1995) Artificial Data Set-B 93

6.7.a Multicollinearity Diagnostics for Hawkins *et. al.* Data..... 110

6.a.1 Simulation Results in Presence of Single High Leverage Point=2.... 114

6.a.2 Simulation Results in Presence of Single High Leverage Point=3.... 114

6.a.3 Simulation Results in Presence of Single High Leverage Point=4.... 115

6.a.4 Simulation Results in Presence of Single High Leverage Point=5.... 115

6.a.5 Simulation Results in Presence of Single High Leverage Point=8.... 116

6.a.6 Simulation Results in Presence of Single High Leverage Point=10. 116

6.b.1 Simulation Results in Presence of 10% Equal High Leverage Point=2... 118

6.b.2 Simulation Results in Presence of 10% Equal High Leverage Point =3. 118

6.b.3 Simulation Results in Presence of 10% Equal High Leverage Point =4. 119

6.b.4 Simulation Results in Presence of 10% Equal High Leverage Point =5. 119

6.b.5 Simulation Results in Presence of 10% Equal High Leverage Point =8. 120

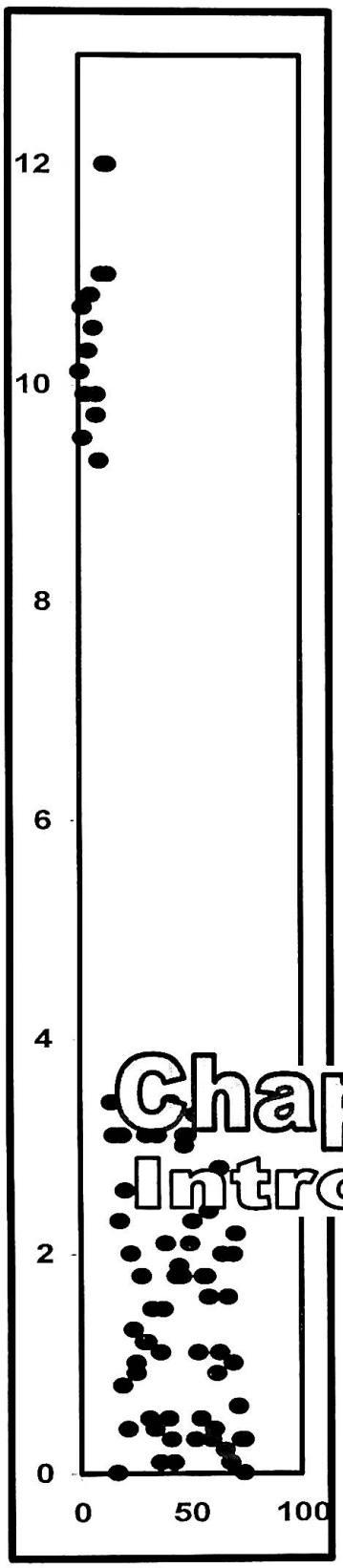
6.b.6 Simulation Results in Presence of 10% Equal High Leverage Point =10. 120

6.c.1 Simulation Results in Presence of 10% Unequal High Leverage Point (2, 4, 6, 8, 10, ..., 40) 122

List of Figures

No	Title	Page No.
2.1	Scatter Diagram that shows Outlier.....	19
2.2	Scatter Diagram that Shows Influential Observation.....	20
2.3	Scatter Diagram that Shows Influential High Leverage Point.....	21
2.4	Scatter Diagram that Distinguishes Outliers, High Leverage Points, and Influential Observations.....	23
2.5	Scatter Diagram that Shows Inter Relation among Outliers, High Leverage Points, and Influential Observations.....	24
3.1	Index Plot of Leverages and Potentials	45
5.1.a	Index Plot of X_1	74

5.1.b	Index Plot of X_2	75
5.1.c	Index Plot of X_3	75
5.2.a	Index Plot of Leverages	75
5.2.b	Index Plot of Potentials	76
5.2.c	Index Plot of Generalizes Potentials	76
5.3.a	L-R Plot for Hawkins *et. al.* (1984) Data	92
5.3.b	P-R Plot for Hawkins *et. al.* (1984) Data	92
5.3.c	GP-DR Plot for Hawkins *et. al.* (1984) Data	92
5.4.a	L-R Plot for Peña and Yohai (1995) Data Set-B	94
5.4.b	P-R Plot for Peña and Yohai (1995) Data Set-B	94
5.4.c	GP-DR Plot for Peña and Yohai (1995) Data Set-B	95
6.a	3D Plot of the Original X's of Hawkins *et. al.* Data	111
6.b	3D Plot of the X's after Deleting the Cases by 2M Method for Hawkins *et. al.* Data	111
6.c	3D Plot of the X's after Deleting the Cases by GP Method for Hawkins *et. al.* Data	112



Chapter One
Introduction

Chapter One

Introduction

The role of high leverage points in linear regression has drawn a great deal of attention in recent years. These points individually or together with some other measures often form the basis of effective diagnostics. Ordinary Least Squares (OLS) technique has been generally adopted in the fitting of regression model because of tradition and ease of computation. We often observe that

> *"not all observations have an equal importance in least squares regression and hence, in conclusions that result from an analysis"* [Chatterjee and Hadi (1986)].

It is, therefore, important to be able to locate such observations and assess their impact on the model. Regression diagnostics mainly deal with cases, which are affected by departures from the assumed model. The prime objective of diagnostic methods is the detection of outliers. In a regression problem, observations corresponding to excessively large random disturbances are treated as outliers. Since the true disturbances are unobserved, they are often estimated by the OLS residuals, which can depend strongly on points in the space of the explanatory

variables which are known as leverage points. Peña and Yohai (1995) point out that high leverage points are mainly responsible for causing masking and swamping of outliers which make the identification procedure of outliers very complicated. Sometimes high leverage points also become responsible for generating the multicollinearity problem in regression analysis. So assessment of high leverage points is equally or sometimes even more important as the identification of outliers in a regression analysis.

A large body of literature is now available for the identification of a single high leverage point and it is generally believed that this problem has already been resolved [see Hawkins, Bradu and Kass (1984)] by the existing methods. But these methods may be ineffective when multiple high leverage points are present in the data.

In chapter two, we discussed in detail about regression, the most popular regression technique, the OLS, and regression diagnostics. We also emphasis upon the discussion about unusual observations such as oultiers, influential observations and high leverage points. In section 2.8, we showed on the basis of some examples that how they differ from one another and the interrelationship among them. We also discussed the consequences of the presence of such unusual observations. Finally we discussed in a brief about robust regression techniques.

In chapter three, we discussed different measures of leverages. Mainly the diagonal elements of the weight matrix are considered as leverage values which measure influences in the X-space, therefore, different properties of the weight matrix (or leverage matrix) and the measures that are used to identify the high leverage points are discussed in this chapter. We also introduced single case-deleted potentials as measures of leverages. The two well known effects, masking

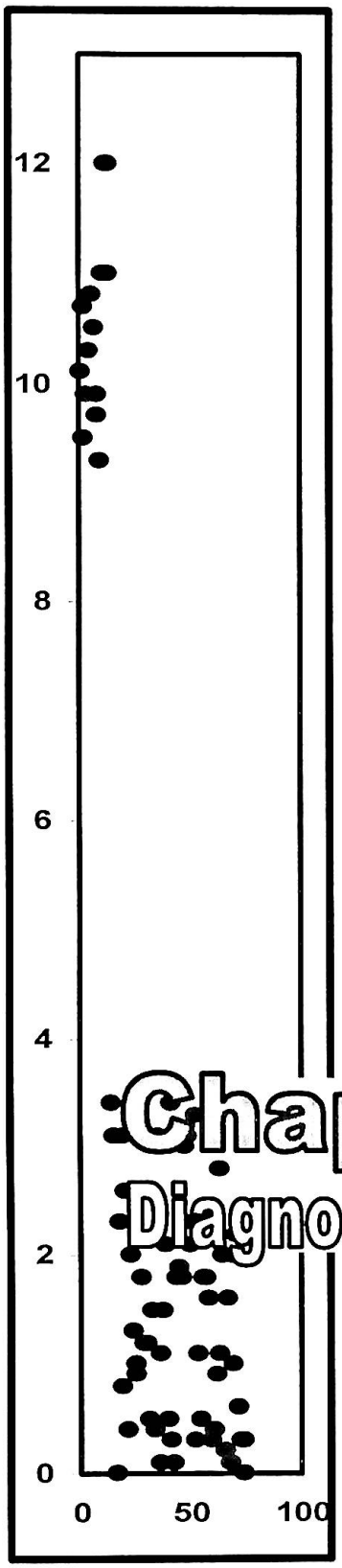
and swamping for which the identification of high leverage points (outliers as well) becomes complicated are also discussed in this chapter.

The performances of different measures of leverages are investigated through a Monte Carlo (MC) simulation in chapter four. At first we briefly discussed about the MC simulation methods. We then considered several examples where different leverage measures are used in both no high leverage and a single high leverage situation. We termed the tendency of any leverage measure to identify cases as high leverage points in a no high leverage situation as sensitivity of that particular measure and compare their sensitivity through an MC simulation. We investigated the usefulness of different leverage measures in the identification of a single high leverage point.

In chapter five, we proposed a new method of detecting multiple high leverage points in linear regression using generalized potential and reported another simulation study which is designed to investigate the performances of the newly introduced method compare to the other existing various commonly used leverage measures in the presence of multiple high leverage points through the Monte Carlo simulation study. We considered both the cases where high leverage points have equal and different weights. After that we also introduced a new graphical display for locating multiple high leverage points together with outliers and influential observations and investigated the performance of this new diagnostic plot along with other existing diagnostic plots.

In chapter six, we reported another Monte Carlo simulation study which is designed to investigate how high leverage points behave as a source of multicollinearity. At first we discussed in a brief what we really mean by multicollinearity and noted its sources, consequences and detection techniques.

We observed how a single high leverage point causes multicollinearity. We also investigated the behavior of multicollinearity when high leverage points thus identified by the existing detection techniques along with generalized potential and omitted from the regression model. We extended this experiment to the case when a group of high leverage points of equal and unequal weights are present in the explanatory variables.



Chapter Two

Diagnostics in Linear Regression

Chapter Two

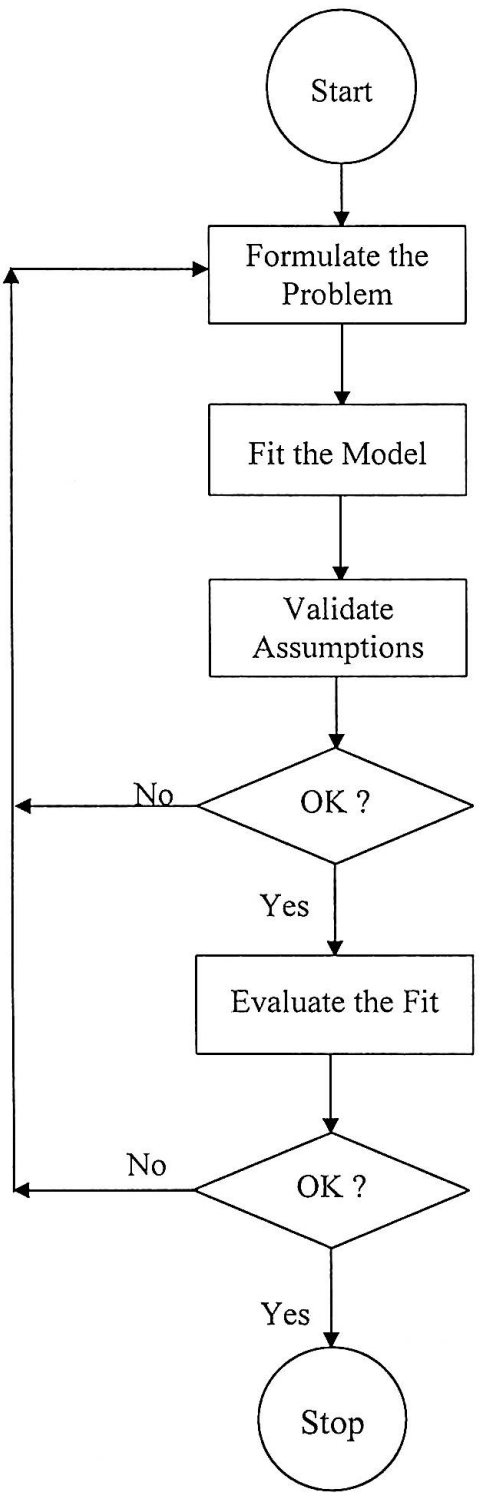
Diagnostics in Linear Regression

Applications of regression are numerous and occur in almost every field, including engineering, the physical sciences, economics, management, life and biological sciences and social sciences. In fact, regression analysis may be the most widely used statistical technique. In this chapter we discussed in detail about regression analysis, the unusual factors those are responsible for the poor fitting of regression model, regression diagnostics and robust regression techniques.

2.1 Regression Analysis

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. In other words, regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variable(s) with a view to estimate and or predict the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) value of the latter.

Chatterjee and Hadi, (1988) have proposed a flow chart as an illustration of iterative regression procedure which is given below:



1. Choose an appropriate set of variables.
2. Choose form of model.
3. Choose method of fitting (e.g. OLS, robust regression, etc.)
4. Specify assumptions.

Using least squares method.

1. Residual plots.
2. Influential data analysis.
3. Test for outliers, etc.

Tests for goodness of fit.

Use the model for the intended purpose.

2.2 Historical Origin of the Term 'Regression'

The term "*Regression*" was introduced by Sir Francis Galton (1886). In a famous paper, Sir Galton reported his findings:

> *"It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species..... It appeared from these experiments that the offspring did not tend to resemble their parent seeds in size, but to be always more mediocre than they to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small.....*
>
> *The experiments showed further that the filial regression towards mediocrity was directly proportional to the parental deviation from it."*

Galton called this phenomenon "*regression toward mediocrity*", later "*mediocrity*" replaced with "*mean*".

Galton's law of universal regression was confirmed by his friends Karl Pearson and A. Lee (1903). They found that the average height of sons of a group of short fathers was greater then their father's height. Thus "*regressing*" tall or short sons alike toward the average height of all men. In the words of Galton, this was "*regression to mediocrity*".

2.3 The Most Popular Regression Technique

Out of many possible regression techniques, the ordinary least squares (OLS) method is the most popular regression technique that was introduced by Carl Friedrich Gauss, a German mathematician. The popularity of ordinary least squares method is attributable to its low computational costs, its intuitive

plausibility in a wide variety of circumstances and its support by a broad and sophisticated body of statistical inference.

Given the data, the tool of least squares can be employed on at least three separate conceptual levels.

First, it can be applied mechanically or descriptively, merely as a means of curve fitting.

Second, it provides a vehicle for hypothesis testing.

Third and most generally, it provides an environment in which statistical theory; discipline specific theory and data may be brought together.

From each of these perspectives, it is often the case that the relevant statistical theory has been quite well developed and those practical guidelines have arisen that make the use and interpretation of least squares straightforward.

2.4 Principles of Ordinary Least Squares Method

We define the simplest form of a general linear model by

$$y_i = x_i^T \beta + \in_i \tag{2.1}$$

Where yi is the i-th observed response,

x_i is a $k \times 1$ vector of predictors,

β is a $k \times 1$ vector of unknown finite parameters and

$\in_i$'s are uncorrelated random errors.

Writing $Y=[y_1, y_2, \cdots, y_n]^T$, $X=[x_1, x_2, \cdots, x_n]^T$ and $\in=[\in_1, \in_2, \ldots, \in_n]^T$ the equation (2.1) can be written as

$$Y=X\beta+\in \tag{2.2}$$

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) possibly including one constant predictor,

β is a $(k\times 1)$ vector of unknown finite parameters to be estimated and

$\in$ is an $n\times 1$ vector of random disturbances.

The assumptions on which several of the least squares results are based are given below:

(a). Linearity Assumption

This assumption is implicit in the defined model (2.1), which says that each observed response value y_i can be written as a linear function of the i-th row of X, x_i^T, that is,

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

(b). Computational Assumption

In order to find a unique estimate of β it is necessary that $(X^TX)^{-1}$ exist or equivalently,

$$\text{rank } (X) = k.$$

(c). Distributional Assumption

The statistical analysis based on least squares (i.e., t-test, F-test etc.) assume that

(i) X is measured without errors,

(ii) $\in_i$ does not depend on x_i^T ; $i=1,2,\cdots,n$ and

(iii) $\in \sim N_n(0,\sigma^2.I)$

(d). The Implicit Assumption

All observations are equally reliable and should have an equal role in determining the least squares results and influencing conclusions.

2.4.1 The Ordinary Least Square Estimator of the Regression Coefficient β

If those assumptions hold, then the OLS estimator of the regression coefficient, β is obtained by minimizing

$$S(\beta) = \epsilon^T \epsilon = (Y - X\beta)^T (Y - X\beta) \tag{2.3}$$

Which implies to,

$$X^T X\hat{\beta} = X^T Y \tag{2.4}$$

Equations (2.4) are the least squares normal equations.

Now premultiplying both sides of (2.4) by $(X^T X)^{-1}$ gives the least squares estimator of β, that is,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.5}$$

The properties of the ordinary least squares estimator, $\hat{\beta}$ are

(i) $\hat{\beta}$ is an unbiased estimator of β, that is, $E(\hat{\beta}) = \beta$.

(ii) $\hat{\beta}$ is the best linear unbiased estimator (BLUE) for β, that is, among the class of linear unbiased estimators, $\hat{\beta}$ has the smallest variance. The variance of $\hat{\beta}$ is

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

(iii) $\hat{\beta} \sim N_k \left[\beta, \sigma^2 (X^T X)^{-1}\right]$.

Where $N_k[\mu,\Sigma]$ denotes a k-dimensional multivariate normal distribution with mean vector μ (a $k\times1$ vector) and variance covariance matrix Σ (a $k\times k$ matrix).

The $n\times1$ vector of fitted (predicted) values of Y is

$$\hat{Y} = X\left(X^TX\right)^{-1}X^TY = WY$$

$$\Rightarrow \hat{Y} = WY \tag{2.6}$$

Where W is the weight matrix, which is defined as

$$W = X\left(X^TX\right)^{-1}X^T \tag{2.7}$$

The following are the properties of the fitted values

(a) $E\left(\hat{Y}\right) = X\beta$

(b) $Var\left(\hat{Y}\right) = \sigma^2W$ and

(c) $\hat{Y} \sim N_n\left[X\beta,\sigma^2W\right]$.

2.5 The Ordinary Least Squares (OLS) Residuals

In regression analysis, since the random errors are unobserved, they are traditionally estimated by the Ordinary Least Squares Residuals, which are actually the differences between observed and estimated responses, when the OLS method is used to fit the model. Mathematically the i-th residual is given by

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

$$\Rightarrow \hat{\epsilon}_i = y_i - x_i^T\hat{\beta}\ ; \qquad i = 1,2,\cdots,n \tag{2.8}$$

In matrix notation

$$\hat{\epsilon} = Y - X^T\hat{\beta} \tag{2.9}$$

$$\Rightarrow \hat{\epsilon} = Y - X^T\left(X^TX\right)^{-1}X^TY$$

$$\Rightarrow \hat{\epsilon} = \left(I - W\right)Y \tag{2.10}$$

That is the ordinary least squares residuals can be expressed in terms of predicted values, Y.

Again we have from equation (2.10)

$$\hat{\epsilon} = (I - W)Y$$
$$\Rightarrow \hat{\epsilon} = (I - W)(X\beta + \epsilon)$$
$$\Rightarrow \hat{\epsilon} = X\beta - WX\beta + (I - W)\epsilon$$
$$\Rightarrow \hat{\epsilon} = (I - W)\epsilon$$
$$\text{i.e } \hat{\epsilon} = (I - W)\epsilon \qquad (2.11)$$

That implies the ordinary least squares residuals can be expressed interms of unobserved errors, ϵ

The ordinary least squares residuals have several properties, which are extremely useful in defining estimation and test procedure based on them. Here we present some of them.

(i) The sum of the residuals in any regression model that contains an intercept term is equal to zero; that is,

$$\sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

(ii) The sum of the residuals weighted by the corresponding value of the predictor always equal to zero; that is,

$$\sum_{i=1}^{n} x_i \hat{\epsilon}_i = 0$$

(iii) The sum of the residuals weighted by the corresponding fitted value is equal to zero; that is,

$$\sum_{i=1}^{n} \hat{y}_i \hat{\epsilon}_i = 0$$

(iv) Under the equation (2.2), the variance of the *i*-th residual is

$$V(\hat{\epsilon}_i) = (1 - w_{ii})\,\sigma^2$$

(v) In this case the covariance between *i*-th and *j*-th residuals is

$$Cov(\hat{\epsilon}_i, \hat{\epsilon}_j) = -w_{ij}\,\sigma^2$$

In both case w_{ii} and w_{ij} are the *i*-th diagonal and *{ij}*-th element of the weight matrix, *W*.

Thus under the equation (2.2) the variance–covariance matrix of the residual vector $\hat{\epsilon}$ can be expressed as

$$Var\ (\hat{\epsilon}) = (I - W)\ \sigma^2$$

(vi) Also the correlation between *i*-th and *j*-th residual is

$$Corr\ (\epsilon_i, \epsilon_j) = \frac{-w_{ii}}{(1 - w_{ii})^{1/2}(1 - w_{jj})^{1/2}}$$

(vii) The residual vector $\hat{\epsilon}$ is distributed as *n*-dimensional multivariate normal distribution with mean vector *0* (zero) and variance–covariance matrix $(I - W)\sigma^2$ i.e,

$$\hat{\epsilon} \sim N_n[\,0, (I - W)\,\sigma^2\,] \text{ and}$$

(vii) $\frac{\hat{\epsilon}^T \hat{\epsilon}}{\sigma^2} \sim \chi^2_{(n-k)}$; where $\chi^2_{(n-k)}$ denotes as χ^2 distribution with *(n-k)* degrees of freedom (*d.f*) and $\hat{\epsilon}^T \hat{\epsilon}$ is the residual sum of squares.

Here an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - k}$$

$$\Rightarrow \hat{\sigma}^2 = \frac{Y^T (I - W) Y}{n - k}$$

2.6 The Weight Matrix, W

The diagonal elements of W, denoted as w_{ii} are called the leverage values which measure how far the input vector x_i are from the rest of the data.

The weight matrix W, which is defined in equation (2.7) plays an important role in least squares regression. The elements of W, denoted by w_{ij} and defined as $w_{ij} = x_i^T (X^T X)^{-1} x_j$, have some nice properties which we shall present in section 3.2.1.

2.7 Regression Diagnostics

It is always necessary to consider how general conclusions would be affected by departures from the assumed model. All of the methods for obtaining estimates and tests based on the OLS technique are computed as if the model and the assumptions are correct, but in many real situations those may be in doubt. So an analysis is designed to check assumptions and build a model is usually required. This is generally known as *regression diagnostics* since they are designed to find problems with assumptions in an analysis.

2.7.1 Departure from Classical Assumptions

Unfortunate consequences of departure from the simple OLS model have long been suspected by statisticians [see Hampel *et al* (1986)]. Despite this fact, the OLS method has retained its popularity over the years in a hope that slight departure from standard assumptions would not affect inferences too much. It is now evident that this type of departure may have drastic consequences on both estimation of parameters and testing of hypotheses.

To quote Tukey (1960)

> *"A tacit hope in ignoring deviations from the ideal model was that they would not matter; that statistical procedures which were optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope is often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians."*

So a check in model adequacy can often be essential in analysing data and using the model.

2.7.2 The Normality Assumption

The normal distribution is the most commonly used distribution in statistical practice. Earlier it was almost a convention to statisticians that the population of the observations would be Normal. This distribution possesses many nice and attractive properties and many classical estimation and testing procedures have been developed on the basis of a Normal assumption of the observations. In the last hundred years attitudes towards a Normal distribution assumption have varied from one extreme to another. According to K Pearson (1905),

> *"... towards the end of the nineteenth century not all were convinced of the need for curves other than normal,"*

By the middle of this century Geary (1947) made this comment,

> *"Normality is a myth; there never was and never will be a normal distribution."*

This might be an overstatement, but the fact that non-Normal distributions are more prevalent in practice than formerly assumed.

2.7.3 Why Testing for Normality Assumption

In a regression context, there has been some effort to argue that the unobserved errors are Normal based on a *priori* considerations [Judge *et al* (1985)]. The errors are perhaps made up of the sum of a large number of separate influences and the distribution of these sum approaches Normal by virtue of the Central Limit Theorem. Under a Normal assumption the OLS method has many desirable properties in both estimation of parameters and test of hypotheses. But in practice we often deal with data sets which are not Normal in nature. If we knew that the distribution of the errors are not Normal, then a different model and different methods could be used, but in reality the true errors are unknown and the are traditionally estimated by the OLS method. Hence it is essential to observe the robustness of OLS estimation and test procedures, which are designed to be optimal at the Normal model.

2.7.4 Problems for Departure from the Normality Assumption

A serious problem may occur when there is any deviation from the Normality assumption, since many of the OLS estimation and test procedures have been developed on the basis of this assumption. A departure from a Normal assumption, such as variance heterogeneity, can cause a great deal of damage to both the estimation and test procedures based on the OLS method.

Gnanadesikan (1977) point out

> *".... the effects on classical methods of departure from normality are neither clearly or easily understood"*

Nevertheless evidence is available that shows such departures can have unfortunate effects in a variety of situations.

In regression problems, Huber (1973) studied the effects of departure from Normality in estimation. He pointed out that under non-Normality it is difficult to find necessary and sufficient conditions such that all estimates of the form $\alpha = \sum_{i=1}^{k} a_i \hat{\beta}_i$ are asymptotically normal. One may also face the problem of estimating correctly the variance–covariance matrix for $\hat{\beta}$.

In testing hypotheses, the effect of departure from Normality has been investigated by many statisticians. A good review of those investigations is available in Judge et al (1985). When $\hat{\epsilon}$ are not Normally distributed, $\hat{\beta}$ and $\frac{(n-k)\hat{\sigma}^2}{\sigma^2}$ are no longer Normal and Chi-square and consequently the t and F test of β are not generally valid in finite samples. However, they have an asymptotic justification. The size of t and F tests appears fairly robust to deviation form Normality [Pearson and Please (1975)]. This robustness of validity is obviously an attractive property, but it is important to investigate the response of test's power as well as size to departure from Normality. Koenker (1982) pointed out that the power of t and F tests is extremely sensitive to the hypothesized error distribution and may deteriorate very rapidly as the error distribution becomes long-tailed. Arnold (1980) studied the distribution of $m_2(\hat{\epsilon}) = \sum_i \hat{\epsilon}_i^2 \Big/ n$ and showed that the significance level of the usual χ^2 test of the hypothesis $\sigma^2 = \sigma_0^2$ is not even asymptotically valid in the presence of non-Normality.

Furthermore, Bera and Jarque (1982) have found that homoscedasticity and serial independence tests suggested for Normal errors may result in incorrect conclusions under non-Normality. It may be also essential to have proper knowledge of distribution in prediction and in confidence limits of predictions. Most of the standard results of this particular study are based on the Normality assumption and the whole inferential procedure may subject to error if there is a departure from this. In all violation of the Normality assumption may lead to the use of suboptimal estimates, invalid inferential statements and inaccurate predictions.

This non-Normality may occur because of their inherent random structure or because of the presence of Outliers or High Leverage Points or Influential Observations.

2.8 Influential Observations, High Leverage Points and Outliers:

In fitting a linear regression model by the OLS method we often observe that a variety of estimates can be substantially affected by one observation or a few observations. Therefore, it is important to be able to locate such observations and assess their impact on the model. In this section we shall discuss three frequently used concepts, that is, Outliers, High Leverage Points and Influential Observations.

2.8.1 Outliers

In a sample of moderate size taken from a population, it can often appear that a few values are surprisingly far away from the main group. According to Barnett and Lewis (1994),

> *"Observations that, in the opinion of the investigator, stand apart from the bulk of the data have been called outliers".*

In the framework of linear regression, we define an outlier to be an observation for which the fitted residual is large in magnitude compared to the other observations in the data set, that is, observations are judged as outliers on the basis of how unsuccessfully the fitted regression equation is in accommodating them and that is why observations corresponding to excessively large residuals.

For example, in the following figure we shall easily illustrate about an outlier.

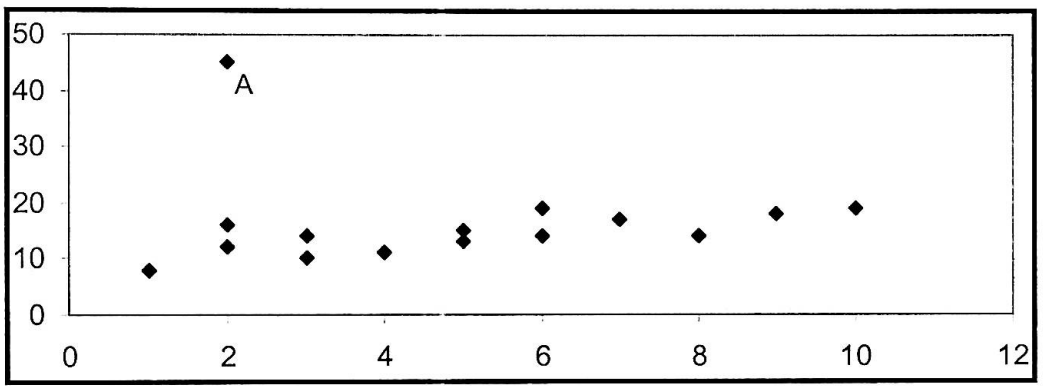


Figure 2.1: Scatter Diagram that Shows Outlier.

In the given figure, if a straight line regression model is fitted to the data, we see clearly that the observation marked by 'A' is an outlier, because it will have a large residual and its omission may not change the slope but will change the intercept of the fitted line. Its omission will also change the estimated error variance, and hence the variance of the estimated coefficients, that is, the fitted line will hardly changed if this point is omitted.

2.8.2 Influential Observations

Influential observations are those observations that, individually or collectively, excessively influence the fitted regression equation as compared to the other observations in the data set. According to Belsley Kuh and Welsch (1980),

> *"An influential observation is one which either individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates then is the case for most of the other observations."*

In this situation, parameter estimates or predictions may depend more on the influential observations than on the majority of the data and their omission from the data may result in substantial changes to important features of an analysis.

If we consider the following example, that is, the following figure, then we shall see that the observation marked by B has a small residual, yet it is omitted, the

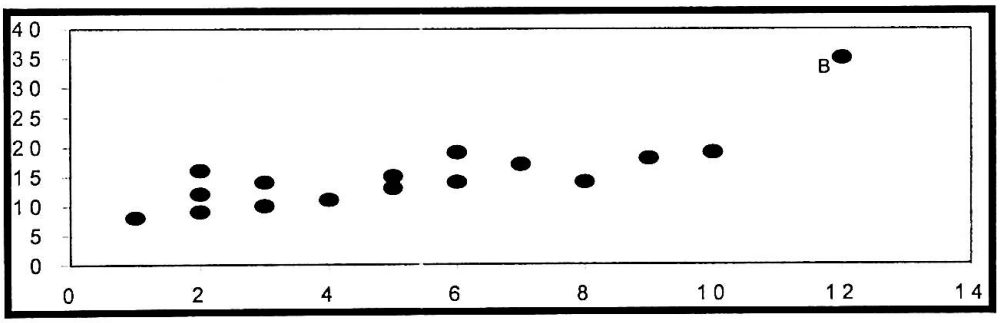


Figure 2.2: Scatter Diagram that Shows Influential Observation.

estimated regression coefficients change substantially. Thus the point marked by 'B' is an example of an influential observation.

2.8.3 High Leverage Points

According to Hocking and Pendletion (1983),

> *"High leverage points are those observations for which the input vector x_i is, in some sense, far from the rest of data."*

Equivalently, a High Leverage Point is an observation with large w_{ii}, the i-th diagonal element of the weight matrix W, in comparison to other observations in the data set. Observations which are isolated in the X space will have high leverage. Points with high leverages may be regard as outliers in X space. The concept of leverage is linked entirely to the predictor variables and not to the response variable.

Consider the following figure as an example, suppose that we currently have the data plotted in the figure. If a straight line regression model is fitted to the data we

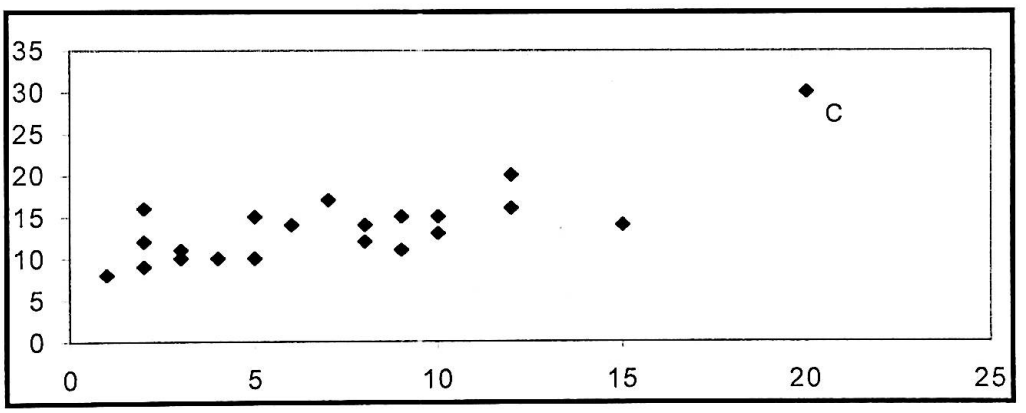


Figure 2.3: Scatter Diagram that Shows High Leverage Point.

shall see that the point marked by 'C' will have a small residual because its Y position is near where the line passes through other points. It will be a High Leverage Point because it is an outlier in X. However, it will not have a large

influence on the fitted regression equation. It is clear that point C is an example of a high leverage point, which is neither an outlier nor an influential point. Also note that point C is influential on the estimated regression coefficients because it is an extreme point in the X space, however, it may be influential on the standard error of the regression coefficients.

2.8.4 Inter relationships among Outliers, Influential Observations and High Leverage Points

It is generally believed that outliers would be highly influential. But that is not always true. Andrews and Pregibon (1978) have presented some examples where outlying observations have little influence on the results. Their examples illustrate the existence of an outlier that does not a matter.

Chatterjee and Hadi (1986) discussed the inter relationship among outliers, influential observations and high leverage points. They observed that,

(a) Influential Observations need not be Outliers in the sense of having high residuals.

(b) Outliers need not be Influential Observations.

(c) While observations with large residuals are not desirable, a small residual does not necessary mean that the corresponding observation is a typical one. This is because least squares fitting avoids large residuals, and thus it may accommodate a point (which is not typical one) at the expense of other points in the data set. In fact there is a general tendency for high leverage points to have small residuals and to influence the fit disproportionately, and

(d) As with outliers, high leverage points need not be influential observations and influential observations are not necessarily high leverage points. However, high leverage points are likely to be influential observations.

To distinguish Outliers, Influential Observations and High Leverage Points we shall consider the following examples:

Example 1: This example illustrating the distinction among outliers, high leverage points and influential observations.

Suppose that we currently have the data plotted in the given figure page and we wish to add one of three points marked by the letters 'A', 'B' and 'C'.

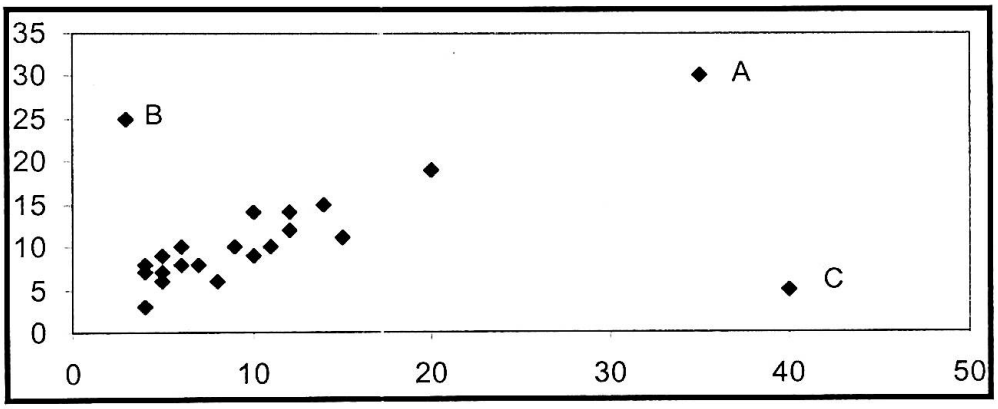


Figure 2.4: Scatter Diagram that Distinguishes Outliers, High Leverage Points and Influential Observations.

If point 'A' is considered for inclusions, it will have a small residual because its Y position is near where the line passes through other points. It will be a high leverage point because it is an outlier in X. However, it will not have a large influence on the fitted regression equation. Hence point 'A' is a high leverage point, which is neither an outlier, nor an influential point.

On the other hand, if point 'B' is considered for addition, it will not be a high leverage point because it is close to the center of X, but it will clearly be an outlier and an influential point. It will have a large residual, and its inclusion may not change the slope but will change the intercept of the fitted line. Its inclusion will also change the estimated error variance, and hence the variance of the estimated coefficients.

Now let us consider the adding point 'C' to the data points. It is clear that point 'C' will be an outlier, a high leverage point, and an influential observation. It will be an outlier because it will have a large residual. It will be a high leverage point because it is an extreme point in the X space, It is an influential observation because its inclusion will substantially change the characteristics of the fitted regression equation.

Example 2: This is an example illustrating that outliers need not be influential observations and influential observations need not be outliers.

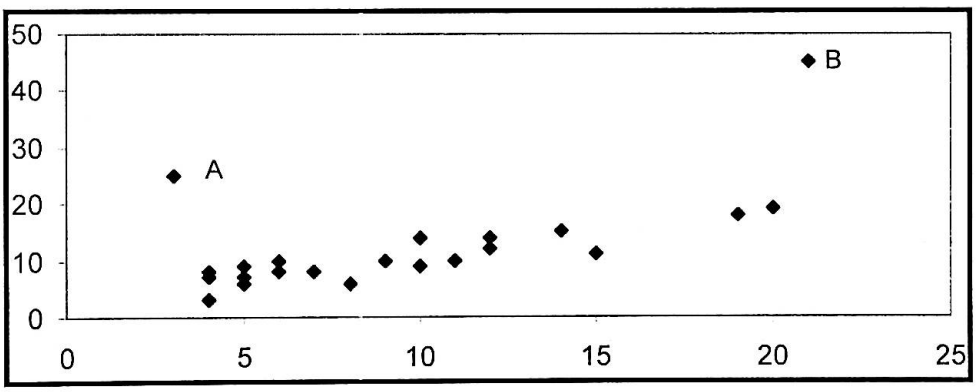


Figure 2.5: Scatter Diagram that Shows Inter Relation among Outliers and Influential Observations.

Consider the data plotted in the given figure. If a straight-line regression model is fitted to the data, we see clearly that the observation marked by 'A' is an outlier.

However, the fitted line will hardly change if this data point is omitted. This is an example of an outlying observation that has a little influence on the estimated regression coefficient. The figure also shows that the observation marked by 'B' has a small residual, yet, when it is omitted, the estimated regression coefficients change substantially. Thus the point marked by 'B' is an example of an influential observation that is not an outlier.

The point marked by 'A' is an outlier, yet the fitted line will hardly change if this point is omitted; whereas the point marked by 'B' has a small residual but is highly influential because of its high leverage. Thus outliers need not be influential observations, and influential observations need not be outliers.

2.8.5 Consequence of the presence of outliers, high leverage points and influential observations

It is generally believed that outlying observations in statistical data are often caused by gross measurement or recording errors. Hampel *et al.* (1986) claim that a routine data set typically contains about 1-10% gross errors, and even the highest quality data set cannot be guaranteed free of gross errors. One immediate consequence of the presence of outlying observations is that they may cause apparent non-Normality and hence one may face the unfortunate consequences mentioned in the previous section. Because the OLS technique minimizes squared deviations, it has a tendency to put a relatively heavy weight on outlying observations and parameter estimates are extremely sensitive to their presence. In OLS method, the residual mean square is generally used to estimate the variance of the errors. The residual mean sum of squares can be greatly inflated by outlying observations so that we may not be able to reliably estimate the variance of the Normal errors.

2.9 Robust Regression Techniques:

In recent years the idea of robustness has been given much more importance in every branch of statistics. To quite Kadane (1984),

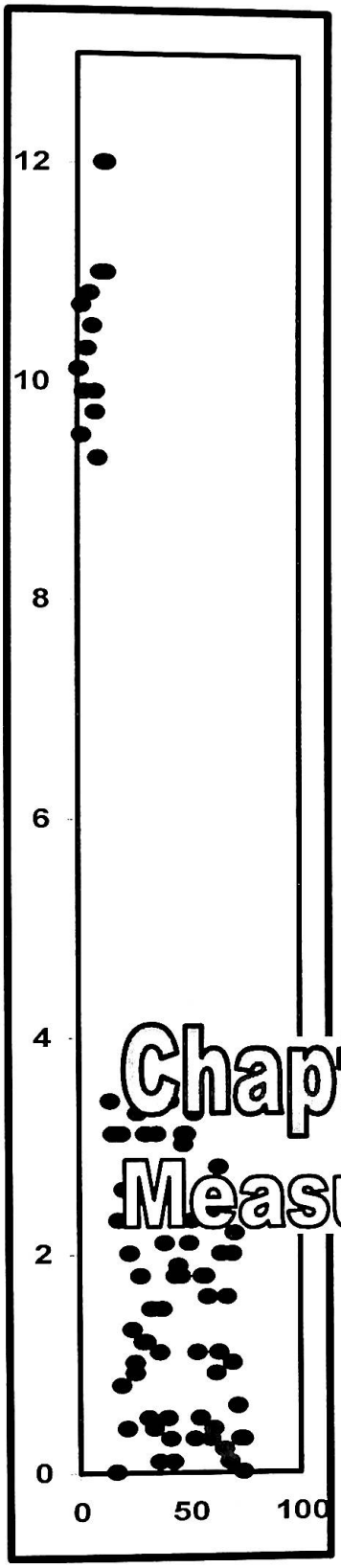
> *"Robustness is a fundamental issue for all statistical analyses; in fact it might be argued that robustness is the subject of statistics".*

One may consider this comment as an overstatement but the importance of robustness cannot be ignored.

In linear regression, robust techniques grew up in parallel to diagnostic [Hampel *et al.* (1986)] and initially they were used to estimate parameters and to construct confidence intervals more efficiently. In recent years diagnostics and robust regression are considered to be complementary to each other [Staudte and Sheather (1990)] in the process of model building and verification. The main application of robust techniques in a regression problem is to try to devise estimators that are not strongly affected by outliers. One objective of robust techniques is to cope with outliers by trying to keep small the effects of their presence. But in recent years, a rationale for this technique has been mainly the identification of multiple outliers. Therefore, diagnostic and robust regression have the same goals, but in the opposite order. To quote Rousseeuw and Leroy (1987),

> *"When using diagnostic tools, one first tries to delete the outliers and then to fit the 'good' data by the least squares, whereas a robust analysis first wants to fit a regression to the majority of the data and then to discover the outliers as those points which possess large residuals from that robust solution."*

In recent years, robust regression techniques are commonly used to identify multiple outliers. Among them 'most likely outlier subset' proposed by Gentleman and Wilk (1975), elemental sets proposed by Hawkins, Bradu and Kass (1984), least median of squares and least trimmed squares proposed by Rousseeuw (1984), reweighted least squares proposed by Rousseeuw and Leroy (1987) have become very popular with the statisticians. In our paper we would suggest a robust procedure to identify multiple high leverage points.



Chapter Three
Measures of Leverages

Chapter Three

Measures of Leverages

In regression analysis, the inferences are highly affected by the outlying observations like as leverages. For accurate inference we must identify these outlying observations otherwise the inference will be inaccurate. But sometimes it is difficult to identify those outlying observations for masking and swamping. In this chapter we shall discuss about Masking and Swamping, the measures of leverages, properties of weight matrix and relation between leverages and potentials.

3.1 Masking and Swamping

As the term masking is most commonly used, it arises when a sample contains multiple outliers but on analysis by a particular outlier detection method, some or all of the outlying observations appear to be inlying. The converse problem of swamping

arises when the method of analysis wrongly suggests that a good data point is outlying. Indeed, two notions of masking have emerged [see Lawrance (1995)], as indicated by the following quotations:

> *"... the structure would not be revealed by the calculation of single deletion diagnostic measures for each observation in turn, although it might well be detected by multiple deletion measures. This effect, which has been called 'masking'..."*[Atkinson (1985)];
>
> *"... there may exist situations, in which observations are jointly but not individually influential, or the other way round ....This situation is sometimes referred to as the masking effect..."* [Chatterjee and Hadi (1988)];
>
> *"...masking effect. This means that after the deletion of one or more influential points, another observation may emerge as extremely influential, which was not visible at first..."* [Rousseeuw and Leroy (1987)];
>
> *"... the importance of a particular observation may not be apparent until some other observation has been delete ... In the presence of such a masking effect..."* [Atkinson (1985)].

The earliest well-known example of masking was given by Pearson and Chandra Sekar (1936). Let $X_1, X_2, \cdots, X_n$ be a sample of size n have mean $\bar{X}$ and variance S^2, and consider the use of the studentized residuals $\frac{(X_i - \bar{X})}{S}$ for outlire detection. Pearson and Chandra Sekar (1936) showed by example that if n is sufficiently small, as X_n and X_{n-1} tend to infinity; the largest studentized residual may tend to a constant below the rejection level. Thus paradoxically as the two outliers become more outlying, the probability of identifying either of them as a significant outlier using the maximum absolute studentized residual goes to zero. This basic

framework also illustrates swamping: fix X_{n-1} and increase X_n until $\bar{X} = X_{n-1}$. Then using the maximum absolute studentized residual, all of the *n-2* good observations appear more outlying than X_{n-1}

In this example the reason for the masking is that the outlying observations inflate *S* by an amount more than compensating for the matching increase in max $(Xi - \bar{X})$. Another much less easily diagnosed problem of masking and swamping can arise in regression, where an additional complication is the leverage of the predictors; i.e., the ability of data point with extreme values of the predictors to lever the regression line over toward themselves. See for example Belsley, *et al* (1980). To illustrate these points, consider a set of (x_i, Y_i) pairs, (20, 20), (10, Δ), (-8, 0) and seven (x_i, Y_i) pairs that are independent $N(0, 1)$. The first two pairs are outliers (if $\Delta \neq 0$); the rest are good. The residuals of the first three points obtained in a simulation with $\Delta = 12$ were 1.30, 1.90, and 5056; using the known $\sigma = 1$, the studentized residuals were 2.25, 2.16, and 6.60. This shows both outliers being masked, with the (-8, 0) inlier being swamped. With $\Delta = 0$, so that the second point is actually inlying, the residuals are 5067, -7.45, and 4.95, and the studentized residuals are 10.58, -8.44, and 5.87. Here the outlier has been unmasked, but the second and third, inlying points, remain swamped.

As the preceding discussion suggests, masking and swamping are deficiencies not of the sample, but of the particular outlier detection method applied to it. For example, while with $\Delta = 12$ the second observation has a larger residual than the first, studentizing to correct for their different variances shows the first to be in fact the more extreme. In the Pearson–Chandra Sekar (1936) case, masking and swamping are easily avoided by replacing $\bar{X}$ and *S* with robust measures of location and scale or by removing the *k* most aberrant points and then successively

testing them for reinclusion. This method, however, will fail in the regression example by swamping the third, good observation pair. In principle the robust estimate remedy applies to the regression, but in practice there may be severe difficulties in finding consistent robust estimators when some points have high leverage. That is why Paña and Yohai (1995) commented that high leverage points are mainly responsible for masking and swamping.

3.2 Measures of Leverages

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. In the past chapter we already defined the high leverage points. Mainly a set of influential X-values is known as a high leverage point. Since residuals are functions of leverage and disturbances (that shown in the past chapter), we observe from (2.11) that high leverage points together with large disturbances (outliers) may pull the fitted least squares line in a way that the fitted residuals corresponding to that outliers might be too small. This may cause masking and/or swamping of outliers [see Peña and Yohai (1995)] and that is why the identification of high leverage points is really necessary.

The i-th diagonal element of the weight $W=X(X^TX)^{-1}X^T$, is traditionally used as measures of leverages.

In the next section we shall discuss different properties of the weight matrix.

3.2.1 Properties of the weight matrix, *W*

The properties of the weight matrix W are given bellow:

Property (i)

W is an idempotent matrix of rank k. By the idempotent property of W it is easy to show that,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^2 = \sum_{i=1}^{n} w_{ii} = k .$$

Proof: We know that if a matrix M is idempotent then $M^TM=M$. Hence

$$\begin{aligned} W^TW &= [X(X^TX)^{-1}X^T]^T[X(X^TX)^{-1}X^T] \\ &= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= W \end{aligned}$$

That is, $W^TW=W$

This implies that the weight matrix, W is an idempotent matrix.

Now we want to find out the rank of W. We know that

$$\begin{aligned} \text{Rank}\,(W) &= tr\,(W) \\ &= tr\,[X(X^TX)^{-1}X^T] \\ &= tr\,[(X^TX)\,(X^TX)^{-1}\,] \\ &= tr\,[I_k] \\ &= k \end{aligned}$$

$\Rightarrow$ Rank $(W)= k.$

Now we shall show that $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^2 = \sum_{i=1}^{n} w_{ii} = k .$

We know by the property of idempotent matrix that

$$W^T W = W, \quad \text{[Since } W \text{ is an idempotent matrix]}$$

$$\Rightarrow \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{n1} \\ w_{12} & w_{22} & \cdots & w_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{nn} \end{bmatrix}^T \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{n1} \\ w_{12} & w_{22} & \cdots & w_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{nn} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{n1} \\ w_{12} & w_{22} & \cdots & w_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{nn} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \sum_{j=1}^{n} w_{1j}^2 & \sum_{j=1}^{n} w_{1j}w_{2j} & \cdots & \sum_{j=1}^{n} w_{1j}w_{nj} \\ \sum_{j=1}^{n} w_{1j}w_{2j} & \sum_{j=1}^{n} w_{2j}^2 & \cdots & \sum_{j=1}^{n} w_{2j}w_{nj} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{j=1}^{n} w_{1j}w_{nj} & \sum_{j=1}^{n} w_{2j}w_{nj} & \cdots & \sum_{j=1}^{n} w_{nj}^2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{1n} \\ w_{12} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{nn} \end{bmatrix} \qquad (3.1)$$

Since we know that

$$tr\,(W) = \text{Rank}\,(W)$$

$$\Rightarrow tr \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{n1} \\ w_{12} & w_{22} & \cdots & w_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1n} & w_{2n} & \cdots & w_{nn} \end{bmatrix} = \text{Rank}\,(W)$$

$$\Rightarrow \sum_{ii=1}^{n} w_{ii} = k \qquad \text{[Since } k \text{ is the rank of } W\text{]} \qquad (3.2)$$

We know that two matrices will be equal iff each element of one is equal to each element of the other.

Hence we get from (3.1) and (3.2)

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^2 = \sum_{i=1}^{n} w_{ii} = k\,. \qquad (3.3)$$

(Proved)

Also we can say from equation (3.2) that

$$\text{Average of } w_{ii} = \frac{k}{n} \tag{3.4}$$

Property (ii)

Here we present further properties of the w_{ii}'s which we use frequently in our subsequent work. For any i and j ranging over $1,2,...,n$

(1) $0 \leq w_{ii} \leq 1$ for all i

(2) $-0.5 \leq w_{ij} \leq 0.5$ [Chatterjee and Hadi (1988)]

(3) If X contains a constant column then for all i, than

$$w_{ii} \geq \frac{1}{n} \quad \text{[Wetherill (1986)]}$$

Proof (ii)(1)

We already got from equation (3.3) that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^2 = \sum_{i=1}^{n} w_{ii}$$

$$\Rightarrow w_{ii} = w_{ii}^2 + \sum_{j\neq i=1}^{n} w_{ij}^2 \tag{3.5}$$

$$\Rightarrow w_{ii} \geq w_{ii}^2 \quad , \text{ Since } \sum_{j\neq i=1}^{n} w_{ij}^2 \geq 0$$

Now since the average of $w_{ii} = \frac{k}{n}$;(n>k) and w_{ii}^2 can not be negative, this implies that

$$0 \leq w_{ii} \leq 1 \quad , \text{ for all } i.$$

(Proved)

Proof (ii)(2)

We get from equation (3.5) that

$$w_{ii} = w_{ii}^2 + \sum_{j \neq i=1}^{n} w_{ij}^2$$

$$\Rightarrow w_{ii} = w_{ii}^2 + w_{ij}^2 + \sum_{r \neq i,j=1}^{n} w_{ir}^2$$

$$\Rightarrow w_{ii} \geq w_{ii}^2 + w_{ij}^2 \quad , \quad Since \quad \sum_{r \neq i,j=1}^{n} w_{ir}^2 \geq 0$$

$$\Rightarrow w_{ij}^2 \leq w_{ii}(1 - w_{ii})$$

Since $0 \leq w_{ii} \leq 1$ this implies that $-0.5 \leq w_{ij} \leq 0.5$

(Proved)

Proof (ii)(3)

If X contains a constant column, define $X=(1 : X)$, where 1 is the n-vector of ones.

Now we know by a property that if $X=(X_1 : X_2)$, where X_1 is an $(n \times r)$ matrix of rank r and X_2 is an $\{ n \times (k- r)\}$ matrix of rank $(k-r)$ and if $W_1 = X_1 (X_1^T X_1)^{-1} X_1^T$ be the prediction for X_1, $V = (I- W_1)X_2$ be the projection of X_2 onto the orthogonal

complement of X_1 . Finally, if $W_2 = V(V^TV)^{-1}V^T$ be the prediction matrix for V, then W (the weight matrix) can be expressed as

$$W = X^T(X^TX)^{-1}X^T = X_1(X_1^TX_1)^{-1}X_1^T + (I - W_1)X_2\{X_2^T(I-W_1)X_2\}^{-1}X_2^T(I - W_1)$$

$$\Rightarrow W = W_1 + W_2 \qquad (3.6)$$

Here $W_1 = 1(1^T1)^{-1}1^T$

$$\Rightarrow W_1 = n^{-1}11^T;$$

Now $V = (I - W_1)X_2 = (I - n^{-1}\mathbf{11}^T)X_2$;

The matrix $(I - n^{-1}11^T)$ is known as centering matrix because it is the linear transformation of X that produces the centered X.

Therefore,

$$W_2 = V(V^TV)^{-1}V^T$$

Thus W can be written as

$$W = W_1 + W_2$$

$$\Rightarrow W = n^{-1}11^T + V(V^TV)^{-1}V^T$$

Now each of the diagonal elements of W_1 is equal to n^{-1} and since W_2 is a prediction matrix, then by property (ii)(2), its diagonal elements are nonnegative, hence

$$w_{ii} \geq n^{-1} = \frac{1}{n}$$

$$\Rightarrow w_{ii} \geq \frac{1}{n} \qquad \text{(Proved)}$$

Accept the above important properties; we shall give some other properties. They are given below:

Property (iii)

For $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, n$

(1) if $w_{ii} = 1$ or 0, then $w_{ij} = 0$.

(2) $(1- w_{ii})(1- w_{jj}) - w_{ij}^2 \geq 0$.

(3) $w_{ii} w_{jj} \geq w_{ij}^2$

(4) $\sum_{j=1}^{n} w_{ij} \hat{\epsilon}_j = 0$ [Montgomery and Peck (1992)]

(5) $w_{ii} + \frac{\hat{\epsilon}_i^2}{\sum_{i=1}^{n} \hat{\epsilon}_i^2} \leq 1$ [Chatterjee and Hadi (1988)]

3.2.2 Different Measures of Leverages

Since in fitting regression model by the Ordinary Least Squares method, estimation and tests are highly affected by the high leverage points, thus we need to identify the leverage points. Much work has been done on the identification of high leverage points and a good number of diagnostic measures are now available in the literature.

In this subsection we shall discuss about different measures of leverages.

(1). Twice the Mean Rule

Hoaglin and Welsch (1978) represented the "*twice the mean rule*" for identification of high leverage points.

We know that the i-th diagonal element, w_{ii} of the weight matrix W is traditionally used measures of leverage of the response value y_i on the corresponding value $\hat{y}_i$. We showed that the average value of w_{ii} is k/n and observations and data points having large w_{ii} values are generally considered as high leverage points. But the immediate questions comes to mind how large is large? Hoaglin and Welsch considered observations are said to be high leverage points when w_{ii} exceeded $2k/n$ and this method is known as "*twice the mean rule*".

(2). Thrice the Mean Rule

Vellman and Welsch (1981) proposed this method for identification of high leverage points. Since data points having large w_{ii} values are generally considered as high leverage points. Vellman and Walsch consider w_{ii} will be large when it exceeds $3k/n$. This method is known as "*thrice the mean rule*".

(3). Huber's Conservative Choice of Cut-off Value for w_{ii}

For a definition of when w_{ii} is large, Huber (1981) suggested to break the range of possible values of w_{ii}. We know that lies between *o* (zero) to *1* (one), that is, $0 \leq w_{ii} \leq 1$ for all i. Huber suggested that w_{ii} will be large when it exceed 0.2, that is, w_{ii} is said to be large if $w_{ii} > 0.2$. This method is known as "*Huber's conservative choice of cut-off value for* w_{ii}", hat we call Huber-1 method.

(4). Huber's Liberal Suggestions

For another definition of when w_{ii} is large, Huber (1981) suggested that w_{ii} is large when it exceed 0.5, that is, w_{ii} is large if $w_{ii} > 0.5$. This method of identification of high leverage points is known as "*Huber's liberal suggestions*", that we call Huber-2 method.

(5). Hadi's Potentials with the Identification Based on Mean

Hadi (1992) pointed out that traditionally used measures of leverages are not sensitive enough to the high leverage points. In the presence of a high leverage point, the weight matrix W may break down easily and after that it may not contain necessary information on high leverage points. In this situation neither of the above methods may be effective in the assessment of the true leverages and consequently the identification of high leverage points becomes complicated. He introduced a new type of measures, named as potentials, where the leverage of the i-th point is based on a fit to the data with the i-th case deleted and that is why is more sensitive to the high leverage points. Every possible subset of $n-1$ observations is used to form the weight matrix, and weight of every deleted observation in turn is generated externally which is known as potentials. Although it seems that calculation of potentials will require construction of n weight matrices, it is possible to calculate them from w_{ii}'s in a very simple way.

Writing the data matrix of k explanatory variables as $X = \left[x_1, x_2, \cdots, x_n\right]^T$, the i-th leverage value is defined as

$$w_{ii} = x_i^T (X^T X)^{-1} x_i$$

We define the i-th potential as

$$p_{ii} = x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i$$

Where $X_{(i)}$ is the data matrix X with the i-th row deleted. Using the result of Miller (1974)

$$\begin{aligned} \left(X_{(i)}^T X_{(i)}\right)^{-1} &= \left(X^T X - x_i x_i^T\right)^{-1} \\ &= \left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} x_i x_i \left(X^T X\right)^{-1}}{1 - x_i^T \left(X^T X\right)^{-1} x_i} \end{aligned} \qquad (3.7)$$

It is easy to obtain a simple relationship between w_{ii} and p_{ii} as

$$p_{ii} = x_i^T (X^T X)^{-1} x_i + \frac{(x_i^T (X^T X) x_i)^2}{1 - x_i^T (X^T X) x_i} = \frac{w_{ii}}{1 - w_{ii}}$$

Observations corresponding to excessively large potential values are considered as high leverage points. Hadi (1992) proposed a cut-off point for p_{ii} as

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

Where c is an appropriately chosen constant such as 2 or 3. This implies that the observations are said to be high leverage points having

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

This form is analogous to a confidence bound for a location parameter. This method is known as " *Hadi's potentials with the identification based on mean*" , which we name *"Potential (mean)"*.

(6). Hadi's Potentials with the Identification Based on Median

In potential (mean) method, the problem with the 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 the alternative suggestion of Hadi (1992) is to consider

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

where the Median Absolute Deviation (MAD) is computed by

$$\text{MAD}\,(p_{ii}) = \text{Median}\,\{\,|p_{ii} - \text{Median}\,(p_{ii})|\,\}$$

and c^* is a suitable chosen constant between 3 and 5.

Hence the observations are said to be high leverage points having

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

This method is known as " *Hadi's potentials with the identification based on median*" , which we name "*Potential (median)*".

(7). Mahalanobis Distance

The leverage of an observation can also be measured by the Mahalanobis distance

Suppose that X contains a column of ones and $\tilde{X}$ denotes the centered X excluding the constant column. A statistic which measures how far x_i is from the center of the data set is commonly computed as $(n-1)^{-1}\tilde{x}_i^T(\tilde{X}^T\tilde{X})^{-1}\tilde{x}_i$, where $\tilde{x}_i$ is the i-th

row of $\tilde{X}$. However, we are interested in measuring how far x_i is from the rest of other observations, hence it is natural to exclude x_i when computing the mean and variance-covariance matrix of X. Therefore, we define Mahalanobis distance as

$$M_i = (n-2)\left\{\tilde{x}_i - \bar{\tilde{X}}_{(i)}\right\}^T \left[\tilde{X}_{(i)}^T \left\{I - (n-1)^{-1} 1\,1^T\right\} \tilde{X}_{(i)}\right]^{-1} \left\{\tilde{x}_i - \bar{\tilde{X}}_{(i)}\right\}, \tag{3.8}$$

Where $\bar{\tilde{X}}_{(i)}$ is the average $\tilde{X}_{(i)}$. Using (3.7) and noting that

$$\bar{\tilde{X}}_{(i)} = (n-1)^{-1} \tilde{X}_{(i)}^T 1 = -(n-1)^{-1} \tilde{x}_i$$

The Mahalanobis distance becomes

$$M_i = \frac{n(n-2)}{n-1} \cdot \frac{w_{ii} - 1/n}{1 - w_{ii}}, \qquad i = 1,2,\cdots,n.$$

3.2.3 Relation Between Mahalanobis Distance and Leverage Values

We can show the relation between Mahalanobis distance and leverage values as follows:

In the case of regression with a constant, let us first split up the x_i into the essential part v_i and the last coordinate 1:

$$x_i = \left(x_{i,1}, x_{i,2}, \cdots, x_{i,k-1}, 1\right) = \left(v_i, 1\right),$$

where $v_i = \left(x_{i,1}, x_{i,2}, \cdots, x_{i,k-1}\right)$ is a $(k-1)$ dimensional row vector.

One computes the mean $\bar{v}$ and covariance matrix C of these v_i:

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$$

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{v})^T(v_i - \bar{v}).$$

One than measures how far v_i from $\bar{v}$ in the matrix define by C, yielding

$$M_i^2 = (v_i - \bar{v})C^{-1}(v_i - \bar{v})^T,$$

which is known as the squared Mahalanobis distance of the i-th case .The purpose of this squared Mahlanobis distance is to point to observations for which the explanatory part lies far from that of the bulk of the data.

Now, we first note that we may subtract the average of each of the first (k-1) explanatory variables, because this changes neither the weight matrix nor the squared Mahalanobis distance. Therefore, we may assume without loss of generality that $\left(\frac{1}{n}\right)\sum_{i=1}^{n} x_{ij} = 0$ for each variable j=1,2,…,(k-1), hence $\bar{v}$=0.

Therefore,

$$X^T X = \begin{bmatrix} x_{11} & \cdots & x_{i1} & \cdots & x_{n1} \\ \vdots & & \vdots & & \vdots \\ 1 & \cdots & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_{11} & \cdots & 1 \\ \vdots & & \vdots \\ x_{i1} & \cdots & 1 \\ \vdots & & \vdots \\ x_{n1} & \cdots & 1 \end{bmatrix}$$

$$\Rightarrow X^T X = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \cdots & \sum_{i=1}^{n} x_{i1}x_{i,k-1} & \cdots & \sum_{i=1}^{n} x_{i1} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{i,k-1}x_{i1} & \cdots & \sum_{i=1}^{n} x_{i,k-1}^2 & \cdots & \sum_{i=1}^{n} x_{i,k-1} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{i,k-1} & \cdots & n \end{bmatrix}$$

$$\Rightarrow X^T X = \begin{bmatrix} & & 0 \\ (n-1)\,\mathbf{C} & & \vdots \\ & & 0 \\ 0 & \cdots & 0 \quad n \end{bmatrix},$$

From which it follows that

$$p_{ii} = x_i \left(X^T X\right)^{-1} x_i^T$$

$$= (v_i \; 1) \begin{bmatrix} \{1/(n-1)\}C^{-1} & 0 \\ 0 & 1/n \end{bmatrix} \begin{bmatrix} v_i \\ 1 \end{bmatrix}$$

$$\Rightarrow p_{ii} = \frac{1}{(n-1)} M_i^2 + \frac{1}{n}$$

This implies that there is a one-to-one relationship between squared Mahalanobis distance and Leverage values.

3.3 Comparison between Potentials and Leverage values

Hadi pointed out that in the presence of high leverage point, potentials are more sensitive then the traditionally used measures of leverages. Here we consider an example. We slightly modify an artificial data set presented by Paña and Yohai (1995), which is given in Table 3.1.

Table-3.1: Modified Paña and Yohai (1995) Data.

Index	Observations (x_i)	Leverages (w_{ii})	Potentials (p_{ii})
1	1	0.3122	0.4539
2	2	0.2315	0.3013
3	3	0.1700	0.2049
4	4	0.1278	0.1465
5	5	0.1047	0.1170
6	6	0.1009	0.1121
7	7	0.1162	0.1315
8	8	0.1508	0.1776
9	9	0.2046	0.2572
10	12	0.4813	0.9278

For this two variable regression model we have one high leverage case (i.e. point 10). We calculate leverage and potentials for this data that are also presented in Table 3.1.

Using the above data we shall form the following figure:

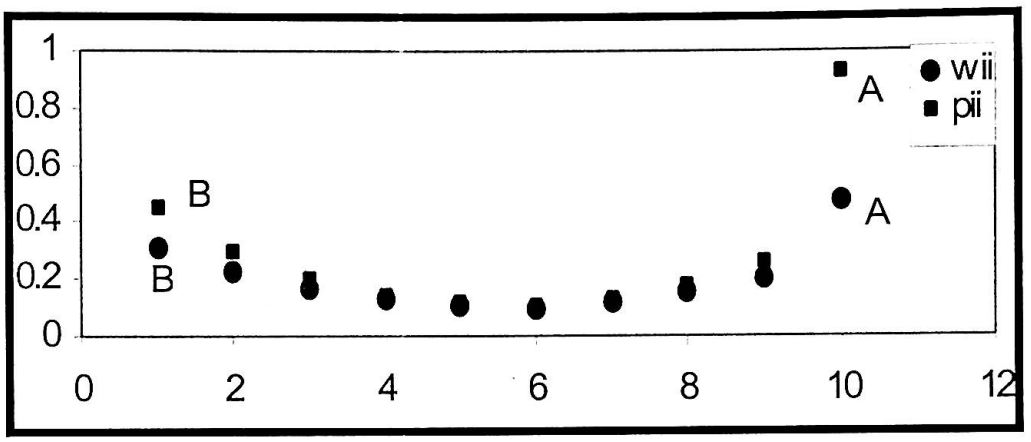
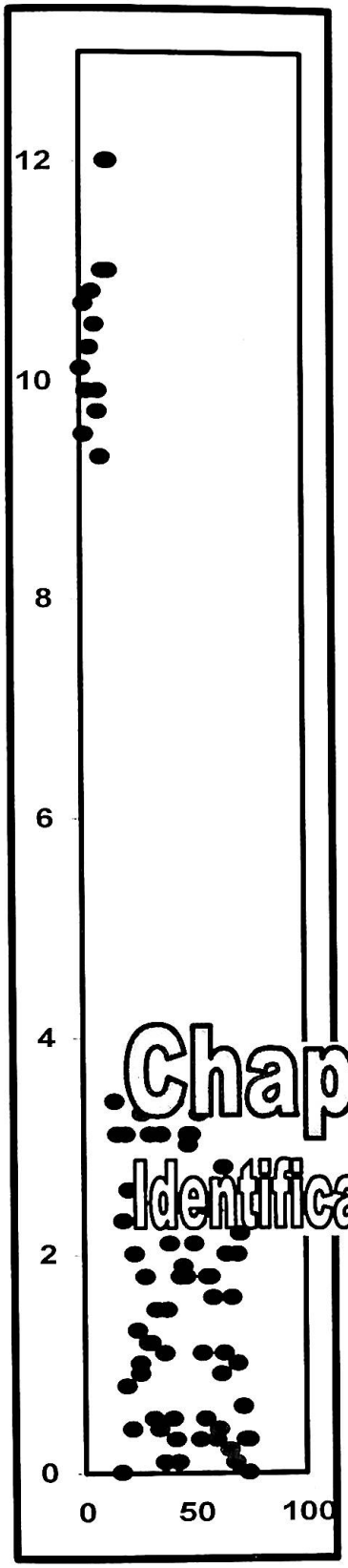


Figure-3.1: Index plot of leverages and potentials.

Figure 3.1 presents the index plot of leverages and potentials for the above data. In this figure the leverage values (w_{ii}) are shown as "•" while the potential values (p_{ii}) are shown as " ■". Values corresponding to the highest leverage and the highest potential values are marked as 'A' and those of the second highest values are marked as 'B'. It is clear from the above figure that for low leverage values, both the leverages and potentials are almost identical. But the potential values are more sensitive for high leverage cases. Even for the second highest leverage value 'B' we observe a marked difference between leverage and potential. This difference is severe for the highest leverage value 'A'. Thus we may conclude that the potential values more sensitive than leverage values.



Chapter Four

Identification of Single High Leverage Point

Chapter Four

Identification of a Single High Leverage Point

In this chapter, we consider several measures of leverages to see how effective they are in the identification of a single high leverage point. We also investigate the sensitivity (in the sense that how this methods identify observations as points of high leverage when in fact there is no high leverage point) of these measures in a no high leverage situation because it is no good if any of the methods identify low leverage cases as points of high leverages. We consider six set of measures, (a) Hoaglin and Welsch's twice-the-mean rule, (b) Vellman and Welsch's thrice-the-mean rule, (c) Huber's conservative choice of cut-off value for w_{ii} (i.e. $w_{ii} > 0.2$), that we call Huber-1, (d) Huber's liberal suggestions (i.e. $w_{ii} > 0.5$), and we call it Huber-2, (e) Hadi's potentials with the identification based on mean $[Mean\ (p_{ii}) + c.St.dev.(p_{ii})]$, which we name Potential (mean) and (f) Potential (median) where the identification criterion is based on $[Median(p_{ii}) + c^*.MAD(p_{ii})]$.

For the last two measures we have chosen c=3 and c^*=5 throughout the experiment.

In our investigation, we consider the Monte Carlo simulation to generate the observations. In next section we shall discuss about simulation.

4.1 SIMULATION

In our research work, we first simulate the X-values keep them fixed to compute all necessary results and then repeat the whole process 10,000 times. Since we use the simulation data in our research work so it is important to know, what is the nature of simulation and how it is performed? In this section, we shall discuss about the basic concept of simulation, and Monte Carlo simulation that is used in research works.

In order to study a system (a system is defined to be the facility or process of interest or a collection of entities, e.g., people or machines that act and interact together toward the accomplishment of some logical end) scientifically we often have to make a set of assumptions about how it works. These assumptions, which usually take the form of mathematical or logical relationships, constitute a model that is used to try to gain some understanding of how the corresponding system behaves.

If the relationship that compose the model are simple enough, it may be possible to use mathematical models (such as algebra, calculus, or probability theory) to obtain exact information on questions of interest; this is called an analytic solution. However, most real-word systems are too complex to allow realistic models to evaluate analytically, and these models must be studied by means of

simulation. In a simulation we use a computer to evaluate a model numerically, and data are gathered in order to estimate the desired true characteristics of the model.

The Monte Carlo Simulation Method is considered as a significant development in the computational statistics. According to Imon (2000), this method helps us to create an artificial "*real type*" situation and the problems from the real world are matched with this simulated situation. In many statistical problems, simulated values are replacing distributional values when no easy way is available to compute them. This method also has been used with several different meanings [Kendal and Buckland (1967)]:

(i) To denote the approximate solution of distributional problems by sampling experiments.

(ii) To denote the solution of mathematical problems arising in a stochastic context by sampling experiments.

(iii) By extension of (ii), the solution of any mathematical problem by sampling methods; the procedure is to construct an artificial stochastic model of the mathematical process and then to perform sampling experiments upon it.

Monte Carlo methods are those in which properties of the distributions of random variables are investigated by use of simulated random numbers. The methods, aside from the collection of data, are similar to the usual statistical methods in which random samples are used in making inference concerning actual populations. Generally in applications of statistics, a model is used to simulate some phenomenon that has a random component. In Monte Carlo methods, on the

other hand, the object of the investigation is a model itself, and random or pseudo-random events are used to study the model.

Often in application of Monte Carlo methods, the problem being studied does not have an explicit random component; however, in these cases a deterministic parameter of the problem is expressed as a parameter of the random distribution and than that distribution is simulated. During World War-*II* and immediately therefore, Monte Carlo methods were extensively used in studying deterministic problems (primarily solutions of differential equations) arising in work on the atomic bomb by Fermi, Von Neumann, Ulam, and Metropolis [see Gentle (1982)]. The name Monte Carlo (from the casino in Monaco) for these methods dates from that period.

Monte Carlo methods are often used by statisticians to investigate distributional problems that are mathematically intractable, such as evaluation of distribution functions or moments of a distribution [Hartley (1977), Imon (1996) and Imon (1999)]. Monte Carlo is also widely used in robustness studies of statistical procedures. The method in this case involves simulating observations from an alternative distribution and computing from these observations the statistics for the procedure in the usual way. From the empirical distributions for the statistics obtained in this manner, the robustness of the ordinary statistical procedure can be evaluated.

Since the introduction of the digital computer, the random numbers used in Monte Carlo studies are most often generated by the computer [Newman and Odell (1971)], and this facility has led to the widespread use of the Monte Carlo technique. In the period from 1978 to 1982, Monte Carlo methods were used in the research is approximately 30% of the articles in the "*Journal of the American*

Statistical Association". The role of Monte Carlo methods has become similar to that of experimentation in the natural sciences, and the need for proper, careful conduct and reporting of Monte Carlo experimentation has been emphasized. A good example of extensive and well-planned use of Monte Carlo is reported in Andrews *et al.* (1972), a study conducted at Princeton University of alternative point estimators of location in symmetric distributions with heavy tails (see Heavy-Tailed Distributions). In such distributions (i.e, Cauchy) the ordinary sample mean is not a good estimator because its variance is large or even infinite. The distributions of many of the other estimators studied are quite intractable, and hence Monte Carlo methods were used to estimate their moments and other properties. Imon (2003) used this type of study to approximate moments and coefficients of skewness and kurtosis of regression residuals. The Monte Carlo studies made extensive use of variance reduction methods, but even so these studies consumed many hours of computer time. While the Monte Carlo results are interesting in their own right, one of the important uses of Monte Carlo is to identify quickly promising statistical methods worthy of further study.

An excellent review of recent computational advances in linear regression and the use of different computer packages in it is available in Ryan (1997) and Imon (2000). Most of the commonly used statistical packages like SPSS, BMDP, MINITAB, SAS, STATA, S-PLUS, LISPSTAT can simulate observations.

Now a day, a large number of computer packages are designed for simulation such as Monte Carlo simulation. Throughout our experiment we use Minitab Version 12.23 for Windows package program for Monte Carlo simulation, since it is simple, readily applicable, safe and very speedy (in the sense of computation time involved in a simulation process).

4.2 Sensitivity of Different Measures of Leverages

To investigate the sensitivity of different measures of leverages let us consider the following examples. The examples are given in the following tables, which show the results for different measures of leverage points in presence of no high leverage point. Here the data sets are generated as Uniform (0,1) for three predictor variables and for different sample sizes such as n=10, 20 and 40 RAND command of MINITAB statistical package for simulating data. Here we have chosen k=4, and c*=5. The six set of measures of leverages are denoted in 2M (twice-the-mean), 3M (thrice-the-mean), Hu1 (Huber's conservative choice), Hu2 (Huber's liberal choice), P.mean (Hadi's potential based on mean) and P.mid. (Hadi's potential based on median).

Example-1

Table-4.a.1: Results of different measure of leverages for n = 10

S.N.	X_1	X_2	X_3	2M	3M	Hu1	Hu2	P.mean	P.med
1	0.864320	0.568475	0.818749	0	0	1	0	0	0
2	0.878396	0.473262	0.618576	0	0	1	0	0	0
3	0.842802	0.303031	0.637115	0	0	1	0	0	0
4	0.682822	0.642951	0.231928	0	0	0	0	0	0
5	0.480566	0.796949	0.272067	0	0	1	0	0	0
6	0.466354	0.845542	0.832563	0	0	1	1	0	1
7	0.397212	0.896034	0.008591	0	0	1	0	0	0
8	0.261917	0.504752	0.392234	0	0	1	1	0	1
9	0.798344	0.354651	0.822990	0	0	1	0	0	0
10	0.869169	0.688198	0.086895	0	0	1	1	0	0

Example-2

Table-4.a.2: Results of different measure of leverages for $n = 20$.

S.N.	X_1	X_2	X_3	2M	3M	Hu1	Hu2	P.mean	Pmed
1	0.113981	0.114947	0.374460	0	0	1	0	0	1
2	0.466643	0.609756	0.747534	0	0	0	0	0	0
3	0.366845	0.638600	0.321566	0	0	0	0	0	0
4	0.942490	0.815849	0.717088	0	0	1	0	0	0
5	0.915512	0.334328	0.571522	0	0	0	0	0	0
6	0.778861	0.194310	0.662487	0	0	0	0	0	0
7	0.594300	0.386302	0.868061	0	0	0	0	0	0
8	0.321238	0.733446	0.657299	0	0	0	0	0	0
9	0.410564	0.119197	0.776130	0	0	0	0	0	0
10	0.243830	0.421769	0.902693	0	0	1	0	0	0
11	0.574727	0.260327	0.151742	0	0	0	0	0	0
12	0.878745	0.514596	0.207859	0	0	0	0	0	0
13	0.761828	0.534407	0.381438	0	0	0	0	0	0
14	0.866674	0.858013	0.380795	0	0	0	0	0	0
15	0.946227	0.381680	0.568873	0	0	0	0	0	0
16	0.842683	0.698638	0.860280	0	0	0	0	0	0
17	0.391032	0.049196	0.023729	0	0	1	0	0	1
18	0.431322	0.811956	0.667877	0	0	0	0	0	0
19	0.349001	0.248340	0.820308	0	0	0	0	0	0
20	0.432600	0.983170	0.053760	1	0	1	0	1	1

Example-3

Table-4.a.3: Results of different measure of leverages for $n = 40$

S.N	X_1	X_2	X_3	2M	3M	Hu1	Hu2	P.mean	P.med
1	0.017070	0.225284	0.640941	0	0	0	0	0	0
2	0.696856	0.867442	0.216194	0	0	0	0	0	0
3	0.040521	0.268515	0.684648	0	0	0	0	0	0
4	0.407694	0.825471	0.248444	0	0	0	0	0	0
5	0.130147	0.053321	0.795865	0	0	0	0	0	0
6	0.055987	0.479750	0.907299	0	0	0	0	0	0
7	0.604971	0.170489	0.667680	0	0	0	0	0	0
8	0.188605	0.101831	0.772676	0	0	0	0	0	0
9	0.384386	0.219363	0.337050	0	0	0	0	0	0
10	0.362117	0.919912	0.991654	0	0	0	0	0	0
11	0.108350	0.161503	0.893317	0	0	0	0	0	0
12	0.942092	0.921921	0.034566	0	0	0	0	0	0
13	0.214357	0.365121	0.704652	0	0	0	0	0	0
14	0.667498	0.801627	0.190790	0	0	0	0	0	0
15	0.706351	0.222835	0.943404	1	0	1	0	1	1
16	0.609853	0.357124	0.543285	0	0	0	0	0	0
17	0.667448	0.981204	0.384749	0	0	0	0	0	0
18	0.232929	0.414669	0.612031	0	0	0	0	0	0
19	0.746899	0.535671	0.505708	0	0	0	0	0	0
20	0.006866	0.286428	0.979625	0	0	0	0	0	0
21	0.263605	0.794207	0.198770	0	0	0	0	0	0
22	0.325486	0.753091	0.882168	0	0	0	0	0	0
23	0.589991	0.171909	0.310800	0	0	0	0	0	0
24	0.192364	0.466240	0.840413	0	0	0	0	0	0
25	0.485929	0.222155	0.098780	0	0	0	0	0	0
26	0.154928	0.369529	0.794789	0	0	0	0	0	0
27	0.147609	0.712685	0.934185	0	0	0	0	0	0
28	0.151357	0.100537	0.768463	0	0	0	0	0	0
29	0.995890	0.992052	0.562380	0	0	0	0	0	0
30	0.167518	0.274811	0.690047	0	0	0	0	0	0
31	0.877638	0.638389	0.307413	0	0	0	0	0	0
32	0.755439	0.872067	0.536830	0	0	0	0	0	0
33	0.166897	0.513302	0.180965	0	0	0	0	0	0
34	0.416015	0.600289	0.285590	0	0	0	0	0	0
35	0.905523	0.833914	0.281912	0	0	0	0	0	0
36	0.471661	0.916765	0.156111	0	0	0	0	0	0
37	0.029969	0.100725	0.535087	0	0	0	0	0	0
38	0.224596	0.243858	0.177413	0	0	0	0	0	0
39	0.567410	0.052890	0.089323	0	0	0	0	0	0
40	0.788180	0.696711	0.073541	0	0	0	0	0	0

4.2.1 Result Discussion for No High Leverage Cases

From Table-4.a.1, we observe that for sample of size 10, the Twice the mean rule, Thrice the mean rule and Potential (mean) rule do not identify any observation as high leverage points but Huber-1 rule, Huber-2 rule and Potential (median) rule identify 9, 3, and 2 observations respectively as high leverage points where as there is no high leverage point in the data sets. Again from Table-4.a.2, we observe that for sample of size 20, the Thrice the mean rule and the Huber-2 rule identify no high leverage point but the Twice the mean rule, the Huber-1 rule, Potential (mean) and Potential (median) rule identify 1, 5, 1, and 3 observations as high leverage point respectively where as there is no high leverage point in the data sets. Similarly from the table-4.a.3 for sample of size 40, we investigate that the Thrice the mean rule and the Huber-2 rule identify no observation as high leverage points but the twice the mean rule, the Huber-1 rule, the potential (mean) and the Potential (median) rule identify 1 (one) observation respectively as a high leverage point where as there is no high leverage point in the data sets.

4.2.2 Simulation Results for Different sample sizes

In this subsection we shall show the results of different commonly used measures of high leverage point after simulating the results 10,000 times by applying the Monte Carlo simulation design.

Table-4.b: Swamping of Cases Using Different Measures of Leverages

Sample size	Measures	Mean		Median	Trmean	Min	Max
n=10	Twice mean	0.118	(1.180)	0.000	0.072	0	2
	Thrice mean	0.000	(0.000)	0.000	0.000	0	0
	Huber-1	9.125	(91.250)	9.000	9.177	5	10
	Huber-2	2.499	(24.990)	3.000	2.497	0	5
	Potential mean	0.000	(0.000)	0.000	0.000	0	0
	Potential Med.	0.800	(8.000)	1.000	0.742	0	4
n=20	Twice mean	0.400	(2.000)	0.000	0.349	0	3
	Thrice mean	0.004	(0.020)	0.000	0.000	0	1
	Huber-1	9.026	(45.130)	9.000	9.007	5	13
	Huber-2	0.040	(0.200)	0.000	0.000	0	1
	Potential mean	0.207	(1.035)	0.000	0.174	0	1
	Potential Med.	0.682	(3.410)	0.000	0.582	0	4
n=30	Twice mean	0.534	(1.780)	0.000	0.467	0	4
	Thrice mean	0.006	(0.020)	0.000	0.000	0	1
	Huber-1	3.440	(11.667)	3.000	3.413	0	8
	Huber-2	0.000	(0.000)	0.000	0.000	0	0
	Potential mean	0.260	(0.867)	0.000	0.231	0	2
	Potential Med.	0.610	(2.033)	0.000	0.498	0	5
n=40	Twice mean	0.644	(1.610)	1.000	0.589	0	3
	Thrice mean	0.008	(0.020)	0.000	0.000	0	1
	Huber-1	0.644	(1.610)	1.000	0.589	0	3
	Huber-2	0.000	(0.000)	0.000	0.000	0	0
	Potential mean	0.310	(0.775)	0.000	0.282	0	2
	Potential Med.	0.484	(1.210)	0.000	0.407	0	4
n=50	Twice mean	0.746	(1.492)	1.000	0.700	0	3
	Thrice mean	0.012	(0.024)	0.000	0.000	0	1
	Huber-1	0.068	(0.136)	0.000	0.020	0	1
	Huber-2	0.000	(0.000)	0.000	0.000	0	0
	Potential mean	0.336	(0.672)	0.000	0.304	0	2
	Potential Med.	0.510	(1.020)	0.000	0.398	0	5
n=100	Twice mean	1.194	(1.194)	1.000	1.113	0	7
	Thrice mean	0.000	(0.000)	0.000	0.000	0	0
	Huber-1	0.000	(0.000)	0.000	0.000	0	0
	Huber-2	0.000	(0.000)	0.000	0.000	0	0
	Potential mean	0.496	(0.496)	0.000	0.433	0	1
	Potential Med.	0.412	(0.412)	0.000	0.320	0	1
n=200	Twice mean	2.054	(1.027)	2.000	1.980	0	7
	Thrice mean	0.000	(0.000)	0.000	0.000	0	0
	Huber-1	0.000	(0.000)	0.000	0.000	0	0
	Huber-2	0.000	(0.000)	0.000	0.000	0	0
	Potential mean	0.674	(0.337)	1.000	0.620	0	4
	Potential Med.	0.298	(0.149)	0.000	0.216	0	3

4.2.3 Simulation Result Discussion for No high leverage Cases

Table-4.b reports a Monte Carlo simulation designed to investigate how sensitive are the different measures of leverages in situations where actually no high leverage points is present. The results of six sets of measures for each of seven samples of size n=10, 20, 30, 40, 50, 100, and 200 are based on the average of 10,000 simulations. This table presents the average results of the mean, the median and the trimmed mean of swamping per sample and the minimum and maximum numbers of observations, which are swamped in each samples for different sample sizes.

This Table clearly shows that Huber 1 method is very sensitive and is not suitable at all for small sample sizes. For a sample of size 10, more than 91% of the total number of observations are appearing as high leverage points. Out of 10 observations, this method identifies on average 9.125 observations as high leverage points with median 9.00 and trimmed mean 9.177. We also observe from the above table that, it identifies minimum 5 observations and maximum 10 observations as high leverage point. Also for a sample of size 20 we observe that, almost half of the total number of observations are appearing as high leverage points, i.e. out of 20 observations, this method identifies on average 9.026 observations as a high leverage points whose median is 9.00 and trimmed mean is 9.007. We also observe from the above table that out of 20 observations, it identifies minimum 5 observations and maximum 13 observations as high leverage points. This swamping rate is over 10% even for n = 30. Some times it identifies 8 observations as high leverage points with mean 3.44, median 3.00 and trimmed mean is 3.413. On the other hand, Huber-2 method is least affected by swamping (except the sample of size n=10 i.e. for sample of size n=10, the swamping rate is

24.99%) followed by thrice-the-mean rule, Potential (mean), twice-the-mean rule and Potential (median). For small sample sizes (e.g. for n=20 and n=30) the last three detection rules have relatively higher (1% to 3%) swamping rates but these rates tend to decrease with the increasing sample sizes. Also for sample of size 10, the Thrice-the-mean rule and the Potential (mean) rule is least affected by swamping.

4.3 Identification of a Single High Leverage Point

In this section we shall investigate how the different measures of leverages are successful in the identification of a single high leverage point when in fact a single high leverage point is present in the data. Throughout the experiment we simulate the first (n-1) observations of the three predictor data are generated as Uniform (0,1) and the n-th observation for each of the three predictors and for each of the sample sizes are fixed at 10, so that the n-th observation will come up as high leverage point. Let us first consider few examples of similar type. We have used the same type of design and same notations as considered in the previous section.

Example-1

Table 4.c.1: Results of different measure of leverages in presence of single leverage point for n = 10.

S.N.	X_1	X_2	X_3	2M	3M	Hu1	Hu2	P.mean	P.med.
1	0.7063	0.0292	0.5548	0	0	1	0	0	0
2	0.2604	0.7856	0.7359	0	0	1	0	0	0
3	0.0893	0.3957	0.5815	0	0	1	0	0	0
4	0.6150	0.7822	0.0353	0	0	1	1	0	0
5	0.2951	0.0423	0.5731	0	0	1	0	0	0
6	0.3893	0.1311	0.5373	0	0	1	0	0	0
7	0.7855	0.4354	0.0269	0	0	1	1	0	0
8	0.3065	0.7322	0.7412	0	0	1	0	0	0
9	0.1071	0.1734	0.4193	0	0	0	0	0	0
10	10.0000	10.0000	10.0000	1	0	1	1	0	1

Example-2

Table 4.c.2 : Results of different measure of leverages in presence of single leverage point for $n = 20$.

S.N.	X_1	X_2	X_3	2M	3M	Hu1	Hu2	P.mean	P.med.
1	0.9665	0.9782	0.4930	0	0	0	0	0	0
2	0.2101	0.2455	0.0753	0	0	0	0	0	0
3	0.7351	0.8326	0.0227	0	0	1	0	0	0
4	0.4299	0.5150	0.7788	0	0	0	0	0	0
5	0.4232	0.2656	0.8255	0	0	0	0	0	0
6	0.4952	0.7402	0.4510	0	0	0	0	0	0
7	0.4704	0.5638	0.9497	0	0	0	0	0	0
8	0.7216	0.4274	0.9457	0	0	0	0	0	0
9	0.0046	0.3272	0.3310	0	0	0	0	0	0
10	0.0657	0.4463	0.7796	0	0	0	0	0	0
11	0.6746	0.2020	0.1134	0	0	1	0	0	0
12	0.3671	0.1818	0.4432	0	0	0	0	0	0
13	0.2491	0.2650	0.1281	0	0	0	0	0	0
14	0.1627	0.9453	0.6101	0	0	1	0	0	0
15	0.7084	0.2945	0.7091	0	0	1	0	0	0
16	0.0140	0.4442	0.9145	0	0	1	0	0	0
17	0.6.55	0.7077	0.6563	0	0	0	0	0	0
18	0.4098	0.3785	0.1028	0	0	0	0	0	0
19	0.7599	0.9111	0.1468	0	0	1	0	0	0
20	10.0000	10.0000	10.0000	1	1	1	1	1	1

Example-3

Table 4.c.3 : Results of different measure of leverages in presence of a single leverage point for n= 40.

S.N.	X_1	X_2	X_3	2M	3M	Hu1	Hu2	P.mean	P.med.
1	0.2530	0.8562	0.2397	0	0	0	0	0	0
2	0.5893	0.4001	0.1474	0	0	0	0	0	0
3	0.7595	0.9799	0.8683	0	0	0	0	0	0
4	0.0469	0.4721	0.2266	0	0	0	0	0	0
5	0.4581	0.2644	0.0290	0	0	0	0	0	0
6	0.1549	0.3776	0.1914	0	0	0	0	0	0
7	0.9884	0.9440	0.0528	0	0	0	0	0	1
8	0.4226	0.6257	0.9384	0	0	0	0	0	0
9	0.2380	0.0844	0.2599	0	0	0	0	0	0
10	0.5719	0.7449	0.9332	0	0	0	0	0	0
11	0.7928	0.5927	0.5169	0	0	0	0	0	0
12	0.8175	0.3623	0.0409	0	0	0	0	0	0
13	0.6792	0.7558	0.1911	0	0	0	0	0	0
14	0.0227	0.8388	0.8965	0	0	0	0	0	0
15	0.3874	0.9378	0.0175	0	0	0	0	0	0
16	0.9622	0.8478	0.9375	0	0	0	0	0	0
17	0.3005	0.5985	0.4179	0	0	0	0	0	0
18	0.5608	0.8601	0.9392	0	0	0	0	0	0
19	0.1415	0.1190	0.8178	0	0	0	0	0	1
20	0.7893	0.9180	0.9082	0	0	0	0	0	0
21	0.2760	0.9202	0.8690	0	0	0	0	0	0
22	0.7558	0.7489	0.3722	0	0	0	0	0	0
23	0.0430	0.9041	0.5897	0	0	0	0	0	0
24	0.3077	0.5388	0.3403	0	0	0	0	0	0
25	0.1018	0.2147	0.6384	0	0	0	0	0	0
26	0.0403	0.9403	0.3576	0	0	0	0	0	0
27	0.4657	0.5788	0.6352	0	0	0	0	0	0
28	0.8509	0.5071	0.0364	0	0	0	0	0	0
29	0.7992	0.9226	0.9942	0	0	0	0	0	0
30	0.9702	0.0959	0.0564	1	0	1	0	0	1
31	0.0825	0.4437	0.5191	0	0	0	0	0	0
32	0.6663	0.2350	0.8165	0	0	0	0	0	0
33	0.2490	0.7195	0.4096	0	0	0	0	0	0
34	0.1599	0.4027	0.4889	0	0	0	0	0	0
35	0.2690	0.6374	0.7086	0	0	0	0	0	0
36	0.3161	0.9621	0.1342	0	0	0	0	0	0
37	0.7508	0.8032	0.9585	0	0	0	0	0	0
38	0.9847	0.5343	0.5861	0	0	0	0	0	0
39	0.8415	0.5043	0.3115	0	0	0	0	0	0
40	10.0000	10.0000	10.0000	1	1	1	1	1	1

4.3.1 Result Discussion for Single High Leverage Case

From Table 4.c.1 we observe that for the sample of size n=10, Huber-1 and Huber-2 rule identify 9 and 3 observations as high leverage points respectively where in fact there is only one high leverage point in the data set. It is also observed from the table that the twice-the-mean rule and potential (median) rule correctly identify the actual high leverage point. But thrice-the-mean rule and potential (mean) rule identify no observations as high leverage points i.e. it is failed to identify the actual high leverage point.

For a sample of size 20, we observe from the table 4.c.2 that the twice the mean rule, the thrice-the-mean rule, Huber-2, potential (mean), potential (median) rules correctly identify one observation as high leverage point. But the Huber-1 rule identifies 6 observations as high leverage points including the actual high leverage point.

We observe from Table 4.c.3 that out of 40 observations with a single high leverage point the thrice-the-mean rule; Huber-2 rule and potential (mean) rule correctly identify the high leverage point. Both the twice-the-mean rule and Huber-1 rule swamp 1 observation each as high leverage point after correctly identifying the actual high leverage point. Also from this table we see that, the potential (median) rule correctly identifies the high leverage point but it swamp three observations.

4.3.2 Simulation Results for Different Measures of Leverages

In this subsection the simulation experiment is designed to investigate how the different measures of leverages are successful in the identification of a single high leverage case. The first (n-1) observations of the three predictor data set for seven sample sizes are generated as Uniform (0,1) and the n-th observation for each of the

three predictors and for each of the sample sizes are fixed at 10, so that the n-th observation will come up as high leverage point in every simulation. The results of this experiment are presented in the following tables (Table 4.d.1–4.d.7) that are based on the average of 10,000 simulations.

Table 4.d.1: Simulation results on the identification of a single high leverage point for $n = 10$

Measures	Identification status	Mean		Median	Tr.mean	Min	Max
2 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.074	(0.74%)	0.000	0.024	0	2
3 Mean	**Identified**	0.000	(0.00%)	0.000	0.000	0	0
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Huber-1	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	6.874	(68.74%)	7.000	0.880	3	9
Huber-2	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	1.448	(14.48%)	0.000	1.440	0	4
Potential (mean)	**Identified**	0.000	(0.00%)	0.000	0.000	0	0
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Potential (median)	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.546	(5.46%)	0.000	0.482	0	3

Table 4.d.2: Simulation results on the identification of a single high leverage point for $n = 20$

Measures	Identification status	Mean		Median	Tr.mean	Min	Max
2 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.194	(0.97%)	0.000	0.151	0	2
3 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.002	(0.01%)	0.000	0.000	0	1
Huber-1	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	5.188	(25.94%)	5.000	5.202	2	8
Huber-2	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.010	(0.05%)	0.000	0.000	0	1
Potential (mean)	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Potential (median)	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.600	(3.00%)	0.000	0.511	0	4

Table 4.d.3: Simulation results on the identification of a single high leverage point for $n = 30$

Measures	Identification status	Mean		Median	Tr.mean	Min	Max
2 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.264	(0.88%)	0.000	0.222	0	2
3 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.002	(0.00%)	0.000	0.000	0	1
Huber-1	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	1.798	(5.99%)	2.000	1.769	0	5
Huber-2	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Potential (mean)	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Potential (median)	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.726	(2.42%)	0.000	0.593	0	6

Table 4.d.4: Simulation results on the identification of a single high leverage point for $n = 40$

Measures	Identification status	Mean		Median	Tr.mean	Min	Max
2 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.238	(0.60%)	0.000	0.178	0	2
3 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Huber-1	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.238	(0.60%)	0.000	0.178	0	2
Huber-2	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Potential	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
(mean)	**Swamped**	0.000	(0.00%)	0.000	0.000	0	0
Potential	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
(median)	**Swamped**	0.704	(1.76%)	0.000	0.576	0	9

Table 4.d.5: Simulation results on the identification of a single high leverage point for $n = 50$

Measures	Identification status	Mean		Median	Tr.mean	Min	Max
2 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.322	(0.64%)	0.000	0.273	0	2
3 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Huber-1	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.028	(0.06%)	0.000	0.000	0	1
Huber-2	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Potential	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
(mean)	**Swamped**	0.000	(0.00%)	0.000	0.000	0	0
Potential	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
(median)	**Swamped**	0.674	(1.39%)	0.000	0.531	0	6

Table 4.d.6: Simulation results on the identification of a single high leverage point for $n = 100$

Measures	Identification status	Mean		Median	Tr.mean	Min	Max
2 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.400	(0.40%)	0.000	0.000	0	4
3 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Huber-1	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Huber-2	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Potential	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
(mean)	**Swamped**	0.000	(0.00%)	0.000	0.000	0	0
Potential	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
(median)	**Swamped**	0.768	(0.77%)	0.000	0.609	0	9

Table 4.d.7: Simulation results on the identification of a single high leverage point for $n = 200$

Measures	Identification status	Mean		Median	Tr.mean	Min	Max
2 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.516	(0.26%)	0.000	0.420	0	5
3 Mean	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Huber-1	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Huber-2	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
	Swamped	0.000	(0.00%)	0.000	0.000	0	0
Potential	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
(mean)	**Swamped**	0.000	(0.00%)	0.000	0.000	0	0
Potential	**Identified**	1.000	(100.00%)	1.000	1.000	1	1
(median)	**Swamped**	0.896	(0.045%)	0.000	0.736	0	7

4.3.3 Simulation Results Discussion for Single High Leverage Case

It is observed from Table-4.d.1 that out of 10 observations twice-the-mean rule correctly identify the leverage value. The success rate of this rule is 100%. The mean, median, and trimmed mean are 1 respectively while the no of high leverage point is also 1. It is also observed from table that, Huber-1, Huber-2 and Potential (median) rules perform similarly for correct identification. But Huber–1 rule swamped 68.74% of the total observations. The mean of swamping is 6.874, median is 7.00 and trim mean is 6.88. This rule identifies minimum 3 and maximum 9 observations as high leverage point, where as there is only one high leverage point present in the data set. The huber-2 and potential (median) rule swamped 14.48% and 5.46% of the total observations. The swamping mean is 1.448 and 0.546, median is 1 and 0, and trim mean is 1.44 and 0.482 respectively. The Huber-2 method swamped minimum 0 and maximum 4 observations. Also the potential (median) rule swamped minimum 0 and maximum 3 observations. On the other hand, the thrice the mean rule and the Potential (mean) rule can not identify any of the high leverage points.

From Table-4.d.2 we observe that out of 20 observations all the six measures successfully identified the high leverage points. The success rate is 100%. But Huber-1 rule swamped 25.94 % of the total observations. The mean, median and trimmed mean of swamping are 5.188, 5.00 and 5.202 respectively. This rule swamped minimum 2 and maximum 8 observations. The potential (median) rule swamped 3% of the total observations. The mean, median and trimmed mean are 0.60, 0.00 and 0.511 respectively. It swamped minimum 0 and maximum 4 observations. The twice-the-mean, thrice-the-mean, and Huber-2 rule swamped 0.97%, 0.10%, and 0.05% of the total observations, the mean is 0.194, 0.002 and

0.005, the median is 0.00 and trim mean is 0.151, 0.00 and 0.00 respectively. The potential (mean) rule swamped no observations.

From Table-4.d.3 we observe that all measures of leverages correctly identify the high leverage points. The success rate is again 100%. The Huber-2 and potential (mean) rule swamped no observations. The twice-the-mean rule swamped 0.88% of the total observations, the swamping mean is 0.264, median is 0.00, trimmed mean is 0.222 respectively and it swamped minimum 0 and maximum 2 observations. The thrice-the-mean rule swamped 0.01% of the total observations. The mean is 0.002, median is 0.00, trimmed mean is 0.00, and it swamped maximum 1 observation. The Huber-1 and potential (median) rule swamped 5.99% and 2.42% of the total observations respectively. The mean, median and trimmed mean is 1.798 and 0.726, 2.00 and 0.00, and, 1.769 and 0.593 respectively. Out of 30 observations, these two rules swamped minimum 0 and maximum 5 and 6 observations respectively.

From the Table-4.d.4 we observe that all the measure of leverages showed a good performance for identification. The rate of identification is 100%. The thrice-the-mean, Huber-2 and potential (mean) swamped no observations. The twice-the-mean, Huber-1 and potential (median) rule swamped 0.60%, 0.60% and 1.76% of the total observations. The mean, median and trimmed mean is (0.238, 0.238 and 0.704), (0.00, 0.00 and 0.00) and (0.178, 0.178 and 0.576) respectively. Out of 40 observations, each of these three methods swamped maximum 2 observations.

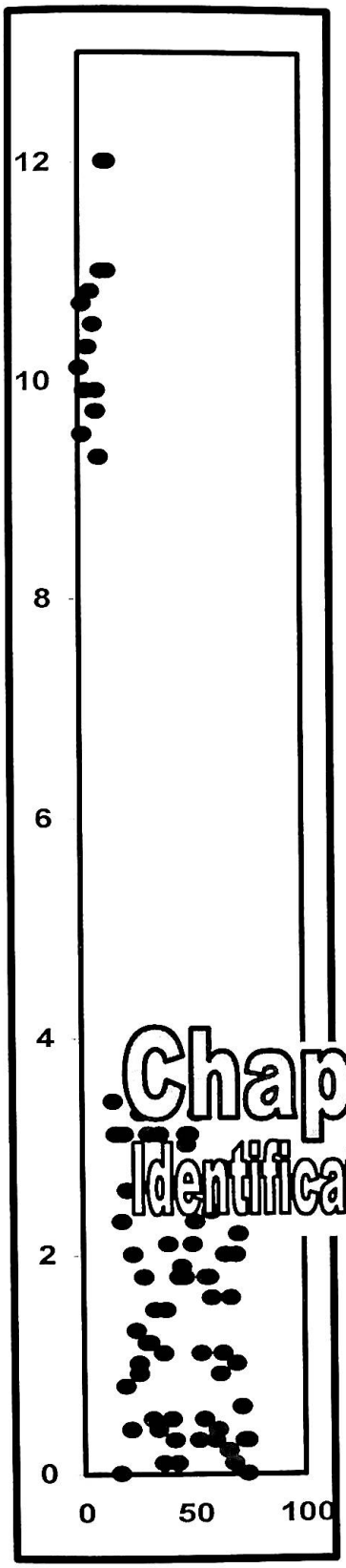
From Table-4.d.5 we observed that all the measure of leverages showed a good performance for identification. The rate of identification is 100%. The thrice-the-mean, the Huber-2 and potential (mean) swamped no observations. The twice-the-mean, Huber-1 and potential (median) rule swamped 0.64%, 0.06% and 1.35% of the total observations. The mean, median and trimmed mean is (0.322, 0.028 and

0.674), (0.00, 0.00 and 0.00) and (0.273, 0.00 and 0.531) respectively. Out of 50 observations, these three methods swamped maximum 2, 1 and 6 observations respectively.

We observe from Table-4.d.5 that all the measure of leverages showed a good performance for identification. The rate of identification is 100%. The thrice-the-mean, the Huber-1, the Huber-2 and potential (mean) swamped no observations. The twice-the-mean, and potential (median) rule swamped 0.40%, and 0.77% of the total observations. The mean, median and trimmed mean is (0.40 and 0.768), (0.00, and 0.00) and (0.316 and 0.609) respectively. Out of 100 observations, these two methods swamped maximum 4 and 9 observations respectively.

Finally from Table-4.d.6 we observe that all the measure of leverages showed a good performance for identification. The rate of identification is 100%. The thrice-the-mean, the Huber-1, the Huber-2 and potential (mean) swamped no observations. The twice-the-mean rule and potential (median) rule swamped 0.26%, and 0.45% of the total observations. The mean, median and trimmed mean is (0.516 and 0.896), (0.00, and 0.00) and (0.420 and 0.736) respectively. Out of 200 observations, these two methods swamped maximum 5 and 7 observations respectively.

From the above discussion we can conclude that, except the sample of size 10, each and every methods considered in this experiment is very successful in the identification of the high leverage point as the success rate is always 100%. But Huber 1 method possesses a very high swamping rate. For a sample of size 20, this rate is over 25% and for $n = 30$, this rate is over 5%. Potential (mean) performed best in this experiment since it did not swamp any of the good cases.



Chapter Five

Identification of Multiple High Leverage Points

Chapter Five

Identification of Multiple High Leverage Points

In the previous chapter we consider cases with no high leverage point or with a single high leverage point. But often we experience that a group of observations in the X-space can exert too much influence in the fitting of a model. This group of observations is known as multiple high leverage points. We observed in chapter four that commonly used leverage measures are being successful in the identification of a single high leverage case. But we anticipate that it is not easy to identify multiple high leverage points. It has been reported by many authors [see Rousseeuw and Leroy (1987), Barnett and Lewis (1994), Peña and Yohai (1995)] that multiple high leverage points are mainly responsible for masking outliers. But the presence of multiple high leverage points may mask themselves in such a way that many of them are not identified when using commonly used leverage measures.

In this chapter we shall propose a new method of detecting multiple high leverage points in linear regression using generalized potential. We shall show how this method works to identify multiple high leverage points compare with the existing methods. At first we present an example and few figures, and then we report a Monte Carlo simulation experiment, which is designed to investigate how, this method along with other six existing methods effective to identify the multiple high leverage points in linear regression.

At last we shall present some graphical techniques, which are also use to identify the multiple high leverage points.

5.1 GENERALISED POTENTIALS

In this section we extend 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. Let us also suppose that R contains $(n\text{-}d)$ cases after $d<(n\text{-}k)$ cases in D are deleted. Without loss of generality, assume that these observations are the last of d rows of X and Y so that the weight matrix $W = X(X^TX)^{-1}X^T$ can be partitioned as

$$W = \begin{bmatrix} U_R & V \\ V^T & U_D \end{bmatrix}$$

Where $U_R = X_R(X^TX)^{-1}X_R{}^T$ and $U_D = X_D(X^TX)^{-1}X_D{}^T$ are symmetric matrices of order $(n\text{-}d)$ and d respectively and $V = X_R(X^TX)^{-1}X_D{}^T$ is a $(n-d)\times d$ matrix. Using the result of Henderson and Searle (1981), $(X_R^TX_R)^{-1}$ can be expressed as

$$(X_R^TX_R)^{-1} = (X^TX)^{-1} + (X^TX)^{-1}X_D^T(I_D - U_D)^{-1}X_D(X^TX)^{-1} \qquad (5.1)$$

Where I_D is an identity matrix of order d. When a group of observations D is omitted, we define

$$w_{ii}^{(-D)} = x_i^T (X_R^T X_R)^{-1} x_i, i = 1,2,\cdots\cdots,n \tag{5.2}$$

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. It can also be expressed by using (5.1) as

$$w_{ii}^{(-D)} = w_{ii} + x_i^T (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} X_D (X^T X)^{-1} x_i$$

Which also implies that for any i,

$$w_{ii}^{(-D)} \geq w_{ii} \tag{5.3}$$

When the size of R is (n-1) and $D = i$, we observe from (2.1) that

$$w_{ii}^{(-i)} = x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i = p_{ii}$$

Which shows that $w_{ii}^{(-D)}$ is a natural extension of p_{ii} .

Suppose now that a further point i is removed from the remaining subset R and joins the deletion subset D. For any such i, from (5.1) and (5.2) it is easy to show that

$$w_{ii}^{-(D+i)} = x_i^T (X_R^T X_R)^{-1} x_i + \frac{(x_i^T (X_R^T X_R)\, x_i)^2}{1 - x_i^T (X_R^T X_R)\, x_i} = \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \tag{5.4}$$

This tells us that the potential value of any case i, generated externally should be equivalent to the quantity $\frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}}$ when $w_{ii}^{(-D)}$ is generated internally on a reduced sample space R. From (5.2) and (5.4) generalised potentials for all members in a data set are defined as

$$p_{ii}^{*} = \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \quad \text{for } i \in R, \text{ for } i \in D \tag{5.5}$$

Where D is any arbitrary deleted set of points. It is obvious from (5.2) and (5.5) that for any i, $p_{ii}^{*} \geq w_{ii}$ and p_{ii}^{*} will be more sensitive to the high leverage points. There exists no finite upper bound for p_{ii}^{*}'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}^{*} to be large if

$$p_{ii}^{*} > Median\,(p_{ii}^{*}) + c^{*}\,MAD\,(p_{ii}^{*}) \tag{5.6}$$

Where $MAD\,(p_{ii}^{*}) = \frac{1}{0.6745}[Median\,\{|\,p_{ii}^{*} - Median\,(p_{ii}^{*})|\}]$

Although the expression of generalised potentials is available for any arbitrary set of deleted cases, D, the choice of such a set is very important. For Hadi's potentials, we have no similar choice. Each and every observation is deleted in turn to determine weights in those cases. But for generalised potentials, it is an important choice which group of observations should be deleted, since the omission of this group determines the weights for the whole set.

Example: The well known Hawkins, Bradu, and Kass (1984) may be a classic example of such a case. They constructed a three predictor artificial data set containing 75 observations with 14 high leverage points (cases 1-14) and 61 low leverage points (cases 15-75) which are given in the next page as Table-5.a.1.

Table 5.a.1: Hawkins-Bradu-Kass (1984) Artificial Data

Index	X_1	X_2	X_3	Y	Index	X_1	X_2	X_3	Y
1	10.1	19.6	28.3	9.7	39	2.1	0.0	1.2	-0.7
2	9.5	20.5	28.9	10.1	40	0.5	2.0	1.2	-0.5
3	10.7	20.2	31.0	10.3	41	3.4	1.6	2.9	-0.1
4	9.9	21.5	31.7	9.5	42	0.3	1.0	2.7	-0.7
5	10.3	21.1	31.1	10.0	43	0.1	3.3	0.9	0.6
6	10.8	20.4	29.2	10.0	44	1.8	0.5	3.2	-0.7
7	10.5	20.9	29.1	10.8	45	1.9	0.1	0.6	-0.5
8	9.9	19.6	28.8	10.3	46	1.8	0.5	3.0	-0.4
9	9.7	20.7	31.0	9.6	47	3.0	0.1	0.8	-0.9
10	9.3	19.7	30.3	9.9	48	3.1	1.6	3.0	0.1
11	11.0	24.0	35.0	-0.2	49	3.1	2.5	1.9	0.9
12	12.0	23.0	37.0	-0.4	50	2.1	2.8	2.9	-0.4
13	12.0	26.0	34.0	0.7	51	2.3	1.5	0.4	0.7
14	11.0	34.0	34.0	0.1	52	3.3	0.6	1.2	-0.5
15	3.4	2.9	2.1	-0.4	53	0.3	0.4	3.3	0.7
16	3.1	2.2	0.3	0.6	54	1.1	3.0	0.3	0.7
17	0.0	1.6	0.2	-0.2	55	0.5	2.4	0.9	0.0
18	2.3	1.6	2.0	0.0	56	1.8	3.2	0.9	0.1
19	0.8	2.9	1.6	0.1	57	1.8	0.7	0.7	0.7
20	3.1	3.4	2.2	0.4	58	2.4	3.4	1.5	-0.1
21	2.6	2.2	1.9	0.9	59	1.6	2.1	3.0	-0.3
22	0.4	3.2	1.9	0.3	60	0.3	1.5	3.3	-0.9
23	2.0	2.3	0.8	-0.8	61	0.4	3.4	3.0	-0.3
24	1.3	2.3	0.5	0.7	62	0.9	0.1	0.3	0.6
25	1.0	0.0	0.4	-0.3	63	1.1	2.7	0.2	-0.3
26	0.9	3.3	2.5	-0.8	64	2.8	3.0	2.9	-0.9
27	3.3	2.5	2.9	-0.7	65	2.0	0.7	2.7	0.6
28	1.8	0.8	2.0	0.3	66	0.2	1.8	0.8	-0.9
29	1.2	0.9	0.8	0.3	67	1.6	2.0	1.2	-0.7
30	1.2	0.7	3.4	-0.3	68	0.1	0.0	1.1	0.6
31	3.1	1.4	1.0	0.0	69	2.0	0.6	0.3	0.2
32	0.5	2.4	0.3	-0.4	70	1.0	2.2	2.9	0.7
33	1.5	3.1	1.5	-0.6	71	2.2	2.5	2.3	0.2
34	0.4	0.0	0.7	-0.7	72	0.6	2.0	1.5	-0.2
35	3.1	2.4	3.0	0.3	73	0.3	1.7	2.2	0.4
36	1.1	2.2	2.7	-1.0	74	0.0	2.2	1.6	-0.9
37	0.1	3.0	2.6	-0.6	75	0.3	0.4	2.6	0.2
38	1.5	1.2	0.2	0.9					

Table 5.a.2: Leverages, Potentials and Generalised Potentials for Hawkins-Bradu-Kass (1984) Data

Index	w_{ii}	p_{ii}	p^*_{ii}	Index	w_{ii}	p_{ii}	p^*_{ii}
1	0.063	0.067	14.46	39	0.035	0.036	0.07
2	0.060	0.064	15.22	40	0.030	0.031	0.04
3	0.086	0.094	16.97	41	0.052	0.055	0.09
4	0.081	0.088	18.02	42	0.055	0.058	0.07
5	0.073	0.079	17.38	43	0.061	0.065	0.09
6	0.076	0.082	15.61	44	0.041	0.043	0.09
7	0.068	0.073	15.71	45	0.029	0.030	0.07
8	0.063	0.067	14.82	46	0.038	0.040	0.07
9	0.080	0.087	17.03	47	0.066	0.071	0.10
10	0.087	0.095	15.97	48	0.041	0.043	0.08
11	0.094	0.104	22.39	49	0.047	0.049	0.06
12	0.144	0.168	24.03	50	0.016	0.016	0.05
13	0.109	0.122	22.73	51	0.036	0.037	0.05
14	0.564	1.294	28.16	52	0.072	0.078	0.09
15	0.058	0.062	0.08	53	0.079	0.086	0.12
16	0.076	0.082	0.09	54	0.040	0.042	0.08
17	0.039	0.041	0.08	55	0.034	0.035	0.05
18	0.023	0.024	0.03	56	0.037	0.039	0.06
19	0.031	0.032	0.04	57	0.023	0.024	0.05
20	0.048	0.050	0.09	58	0.040	0.042	0.07
21	0.029	0.030	0.04	59	0.019	0.019	0.04
22	0.046	0.048	0.07	60	0.062	0.066	0.09
23	0.029	0.030	0.04	61	0.051	0.054	0.10
24	0.026	0.027	0.05	62	0.021	0.021	0.08
25	0.022	0.022	0.08	63	0.036	0.037	0.07
26	0.032	0.033	0.07	64	0.026	0.027	0.07
27	0.042	0.044	0.08	65	0.031	0.032	0.06
28	0.024	0.025	0.03	66	0.036	0.037	0.05
29	0.018	0.018	0.04	67	0.019	0.019	0.02
30	0.047	0.049	0.09	68	0.046	0.048	0.09
31	0.059	0.057	0.07	69	0.029	0.030	0.07
32	0.036	0.037	0.07	70	0.027	0.028	0.05
33	0.026	0.027	0.04	71	0.019	0.019	0.03
34	0.032	0.033	0.09	72	0.028	0.029	0.03
35	0.034	0.035	0.08	73	0.043	0.045	0.05
36	0.023	0.024	0.04	74	0.050	0.053	0.05
37	0.059	0.063	0.08	75	0.062	0.066	0.09
38	0.021	0.021	0.05				

Table-5.a.2 presents the commonly used leverage values w_{ii} together with Hadi's potential values p_{ii} and generalised potential p_{ii}^{*}.

It is clear from the results presented in the Table-5.a.2 that w_{ii} values corresponding to the most of the high leverage points are not large enough and if any one considered Vellman and Welsch (1981)'s "*thrice-the-mean*" rule only observation 14 appears as the point of high leverage. 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 first 13 cases. But generalised potential clearly distinguishes 14 high leverage points from other observations.

Index plot of each of the three regressors of Hawkins-Bradu-Kass (1984) data is presented in figure-5.1 while figure-5.2 presents index plots of leverages, and potentials for the same.

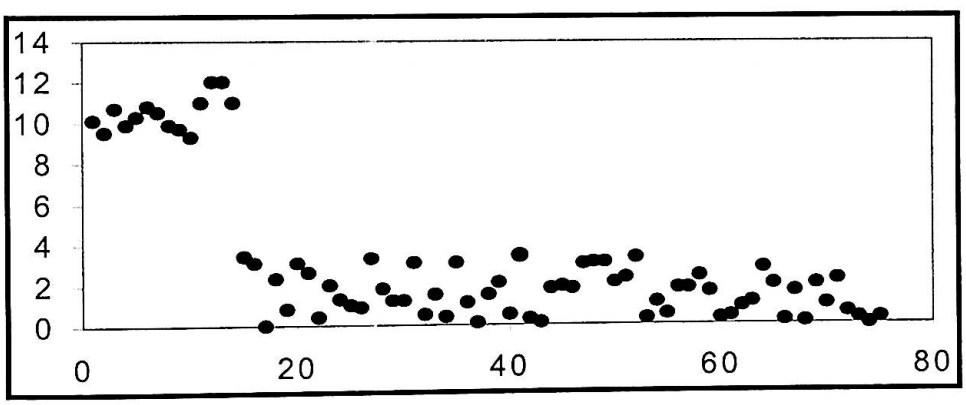


Figure 5.1.a: Index Plot of X_1

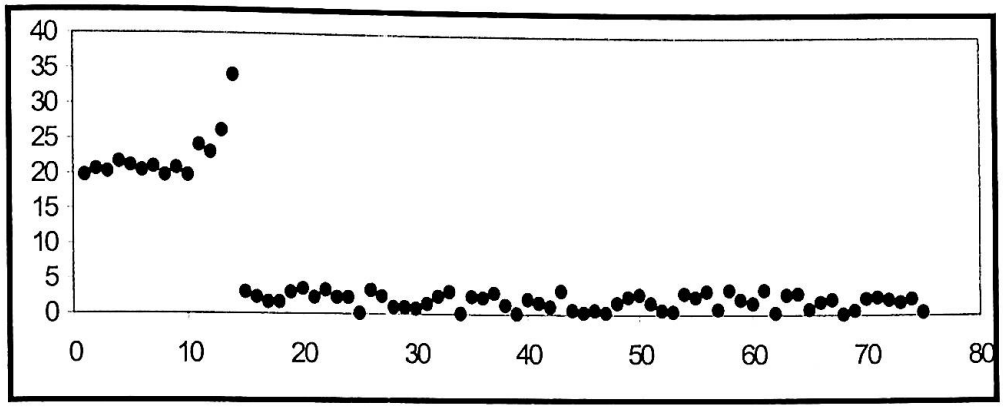


Figure 5.1.b: Index Plot of X_2

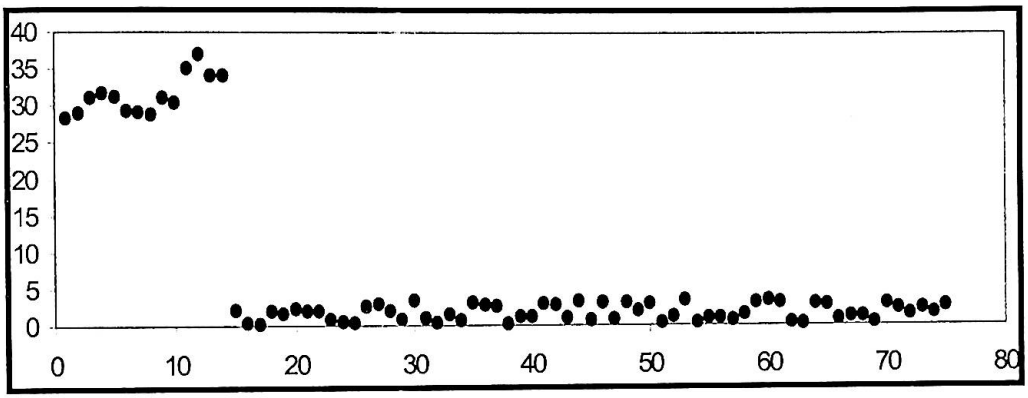


Figure 5.1.c Index Plot of X_3

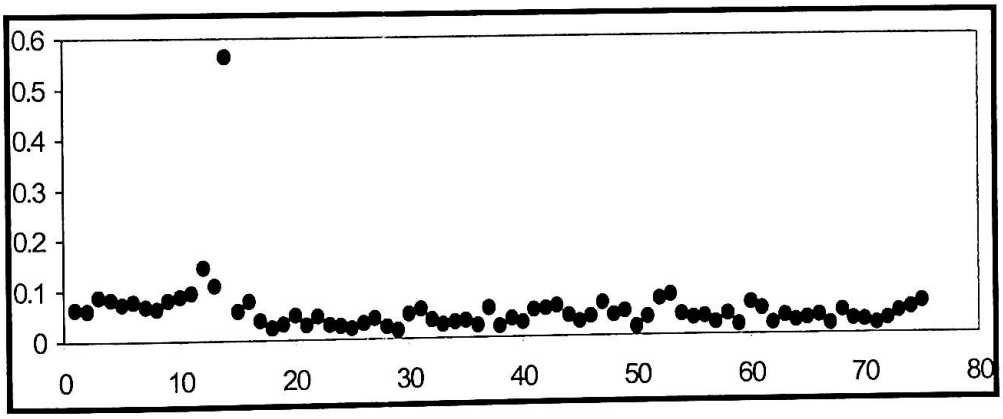


Figure 5.2.a: Index plot of Leverages

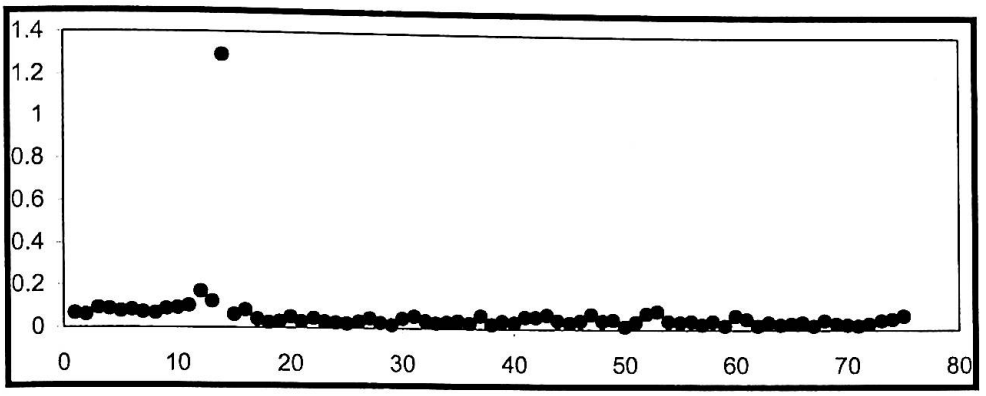


Figure 5.2.b: Index Plot of Potentials

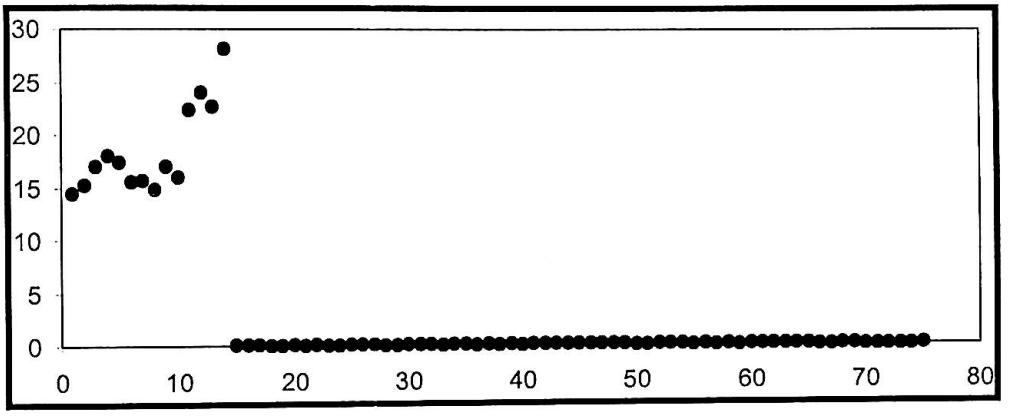


Figure 5.2.c: Index Plot of Generalized Potentials

From the above figures we observe that Figures 5.1.a-5.1.c clearly show the first 14 observations are too far from the rest of the data. It is interesting to see from figures 5.2.a and 5.2.b that when w_{ii} and p_{ii} are considered only case 14 appears as high leverage point in their respective index plots. The other 13 high leverage points appear as points of low leverages. Although these 14 points have similar X

values, w_{ii} and p_{ii} value for the 14-th observation is different from those of the first 13 observations. Thus the first 13 observations are masking from the leverage point of view when the first 14 observations are high leverage points. In order to avoid this problem we introduced "*Generalised Potential*". From figure 5.2.c we clearly observe that the first 14 observations are high leverage points.

5.2 Identification of Multiple (10%) Equally High Leverage Points

In this section we shall investigate how successfully the different measures of leverage identify high leverage cases when a group of leverage points are present in the data.

5.2.1 Different examples for the performance of the seven sets of measures

In order to compare the performances of six sets of measures as considered in the previous chapter for the identification of multiple equally high leverage points, we shall consider at first the cases where high leverage points are of same value. Let us consider the following examples. For the three examples considered here we generate the first 90% observations for each of the three predictor data set as Uniform (0, 1) by using the Monte Carlo simulation design and the last 10% observation for each of the three predictors and for each of the sample sizes are fixed at 10, so that the last 10% observations will come up as high leverage point.

Example-1

Table-5.b.1 Results of measures of leverages for 10% high leverage data for n=20

S.N.	X_1	X_2	X_3	2M	3M	Hu-1	Hu-2	P.mean	P.med	G.P.
1	0.1458	0.6596	0.3083	0	0	0	0	0	0	0
2	0.8435	0.8946	0.1619	0	0	1	0	0	0	0
3	0.6859	0.8649	0.7653	0	0	0	0	0	0	0
4	0.6105	0.2509	0.4001	0	0	0	0	0	0	0
5	0.2787	0.8215	0.2578	0	0	0	0	0	0	0
6	0.5008	0.7062	0.5396	0	0	0	0	0	0	0
7	0.7945	0.7248	0.5502	0	0	0	0	0	0	0
8	0.9151	0.6714	0.3794	0	0	0	0	0	0	0
9	0.1353	0.5854	0.9806	1	0	1	0	0	1	0
10	0.9025	0.1465	0.7231	1	0	1	0	0	1	0
11	0.5578	0.2271	0.1733	0	0	0	0	0	0	0
12	0.2467	0.6634	0.5191	0	0	0	0	0	0	0
13	0.3145	0.9176	0.7040	0	0	0	0	0	0	0
14	0.9918	0.6518	0.4217	0	0	0	0	0	0	0
15	0.0622	0.9769	0.3183	0	0	1	0	0	0	0
16	0.7060	0.6193	0.4252	0	0	0	0	0	0	0
17	0.6484	0.8561	0.8119	0	0	0	0	0	0	0
18	0.8268	0.8738	0.7385	0	0	0	0	0	0	0
19	10.0000	10.0000	10.0000	1	0	1	0	0	1	1
20	10.0000	10.0000	10.0000	1	0	1	0	0	1	1

Example-2

Table-5.b.2 Results of measures of leverages for 10% high leverage data for n=30

S.N.	X_1	X_2	X_3	2M	3M	Hu-1	Hu-2	P.mean	P.med	G.P.
1	0.6608	0.0605	0.8122	0	0	0	0	0	0	0
2	0.3945	0.6917	0.5998	0	0	0	0	0	0	0
3	0.4442	0.8255	0.8999	0	0	0	0	0	0	0
4	0.8128	0.5395	0.5000	0	0	0	0	0	0	0
5	0.1807	0.6104	0.1908	0	0	0	0	0	0	0
6	0.7162	0.4220	0.3534	0	0	0	0	0	0	0
7	0.5662	0.9231	0.5169	0	0	0	0	0	0	0
08	0.2394	0.9454	0.5043	0	0	0	0	0	0	0
9	0.0742	0.3208	0.3096	0	0	0	0	0	0	0
10	0.0728	0.3762	0.9164	0	0	1	0	0	0	0
11	0.0355	0.8962	0.1278	0	0	0	0	0	0	0
12	0.1621	0.8259	0.3893	0	0	0	0	0	0	0
13	0.9170	0.1571	0.9019	0	0	0	0	0	0	0
14	0.5297	0.4689	0.5155	0	0	0	0	0	0	0
15	0.0793	0.4790	0.8532	0	0	0	0	0	0	0
16	0.7856	0.1278	0.8728	0	0	0	0	0	0	0
17	0.2602	0.4841	0.1598	0	0	0	0	0	0	0
18	0.1555	0.3042	0.5285	0	0	0	0	0	0	0
19	0.2943	0.1164	0.4657	0	0	0	0	0	0	0
20	0.4659	0.0266	0.0031	0	0	0	0	0	0	0
21	0.0154	0.1982	0.1588	0	0	0	0	0	0	0
22	0.8779	0.6096	0.7665	0	0	0	0	0	0	0
23	0.7688	0.2673	0.8404	0	0	0	0	0	0	0
24	0.6194	0.4255	0.5032	0	0	0	0	0	0	0
25	0.3685	0.8498	0.4523	0	0	0	0	0	0	0
26	0.9078	0.9798	0.1536	1	0	1	0	0	1	0
27	0.5305	0.4324	0.9982	0	0	0	0	0	0	0
28	10.0000	10.0000	10.0000	1	0	1	0	0	1	1
29	10.0000	10.0000	10.0000	1	0	1	0	0	1	1
30	10.0000	10.0000	10.0000	1	0	1	0	0	1	1

Example-3

Table-5.b.3 Results of measures of leverages for 10% high leverage data for n=40

S.N.	X_1	X_2	X_3	2M	3M	Hu-1	Hu-2	P.mean	P.med	GP.
1	0.3978	0.0315	0.1923	0	0	0	0	0	0	0
2	0.2707	0.7609	0.6883	0	0	0	0	0	0	0
3	0.8935	0.5368	0.1118	0	0	0	0	0	0	0
4	0.7309	0.8714	0.4200	0	0	0	0	0	0	0
5	0.4913	0.2684	0.7213	0	0	0	0	0	0	0
6	0.9929	0.1250	0.9353	0	0	0	0	0	0	0
7	0.1107	0.0912	0.9285	0	0	0	0	0	0	0
8	0.1632	0.6284	0.5632	0	0	0	0	0	0	0
9	0.8475	0.9535	0.8187	0	0	0	0	0	0	0
10	0.5185	0.6364	0.6536	0	0	0	0	0	0	0
11	0.7971	0.5416	0.9988	0	0	0	0	0	0	0
12	0.0184	0.3178	0.5986	0	0	0	0	0	0	0
13	0.0777	0.2884	0.5982	0	0	0	0	0	0	0
14	0.7331	0.2946	0.1659	0	0	0	0	0	0	0
15	0.1451	0.5110	0.8217	0	0	0	0	0	0	0
16	0.6507	0.6811	0.8198	0	0	0	0	0	0	0
17	0.7319	0.2811	0.8801	0	0	0	0	0	0	0
18	0.8869	0.6271	0.2023	0	0	0	0	0	0	0
19	0.7162	0.9966	0.2819	0	0	0	0	0	0	0
20	0.9296	0.4457	0.9908	0	0	0	0	0	0	0
21	0.3103	0.6357	0.2463	0	0	0	0	0	0	0
22	0.0792	0.1556	0.8185	0	0	0	0	0	0	0
23	0.5296	0.4175	0.7495	0	0	0	0	0	0	0
24	0.5708	0.5601	0.7038	0	0	0	0	0	0	0
25	0.3262	0.2012	0.8336	0	0	0	0	0	0	0
26	0.8654	0.7048	0.3735	0	0	0	0	0	0	0
27	0.4372	0.2928	0.6551	0	0	0	0	0	0	0
28	0.1371	0.5677	0.8451	0	0	0	0	0	0	0
29	0.2672	0.5275	0.7127	0	0	0	0	0	0	0
30	0.9564	0.0211	0.5381	1	0	1	0	0	0	0
31	0.3804	0.6439	0.1322	0	0	0	0	0	0	0
32	0.4278	0.3384	0.8871	0	0	0	0	0	0	0
33	0.7952	0.1675	0.1940	0	0	0	0	0	0	0
34	0.4783	0.6712	0.7999	0	0	0	0	0	0	0
35	0.6201	0.7163	0.8226	0	0	0	0	0	0	0
36	0.4760	0.8758	0.2251	0	0	0	0	0	0	0
37	10.0000	10.0000	10.0000	1	0	1	0	0	1	1
38	10.0000	10.0000	10.0000	1	0	1	0	0	1	1
39	10.0000	10.0000	10.0000	1	0	1	0	0	1	1
40	10.0000	10.0000	10.0000	1	0	1	0	0	1	1

5.2.2 Result Discussion

We observe from Example-1 that when multiple (10%) high leverage points are present the twice-the-mean rule, the Huber-1 rule and the potential (median) rule correctly identify the high leverage points. They also identify 2, 4 and 2 observations respectively as high leverage points, which are in fact not high leverage points. Rest of the three rules fail to identify any observations as high leverage points though there are two high leverage points present in the data set. But Generalised Potential correctly identifies the high leverage points and it does not swamp any good observations as high leverage points.

Similar results are obtained found from Example-2. Here twice-the-mean rule, Huber-1 rule and potential (median) rule correctly identify the three high leverage points. They also respectively swamp 1, 2 and 1 other observation(s). Rest of the three rules can't identify any observations as high leverage points i.e., three high leverage points are masked here. Here too Generalised Potential correctly identifies the high leverage points and it does not identify any good observations as high leverage points or masks any high leverage points.

We also observe from example-3 that the twice-the-mean rule, the Huber-1 rule and potential (median) rule correctly identify the four high leverage points. It is also observed that twice-the-mean rule and Huber-1 rule swamped another observation each. In this example too, rest of the three rules cannot identify any observations as high leverage points though there are four high leverage points present in the data set. Generalised Potential produced the best set of results for this data.

5.2.3 Simulation Results

Here we report a Monte Carlo simulation experiment designed to compare the performances of seven sets of measures in the identification of multiple high leverage points when 10% of the data set are of equally high leverage. We constructed the data sets for different sample sizes as it was generated for the examples presented in the subsection 5.2.1. Tables 5.c.1-5.c.6 present simulation results of the measures of equally high leverage points for different sample sizes. Each of the results presented here is based on the average of 10,000 simulations.

Table 5.c.1: Simulation results in presence of multiple (10%) equally high leverage point (2)

Measures		Identification Status	Sample size, *n*					
			20	30	40	50	100	200
2 Mean		Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.268	0.420	0.563	0.605	1.063	1.910
3 Mean		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.000	0.008	0.008	0.010	0.008	0.000
Huber-1		Identified	2.000	3.000	4.000	0.000	0.000	0.000
		Swamped	5.710	2.425	0.563	0.065	0.000	0.000
Huber-2		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.038	0.000	0.000	0.000	0.000	0.000
Potential (mean)		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.008	0.025	0.028	0.020	0.018	0.003
Potential (med.)		Identified	1.860	2.318	2.560	2.963	2.525	1.000
		Swamped	0.470	0.390	0.368	0.323	0.208	0.145
Generalised Potential	c=3	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.208	0.150	0.200	0.170	0.130	0.125
	c=5	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.013	0.010	0.005	0.010	0.005	0.000

Table 5.c.2: Simulation results in presence of multiple (10%) equally high leverage point (3)

Measures		Identification Status	Sample size, *n*					
			20	30	40	50	100	200
2 Mean		Identified	20.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.258	0.398	0.525	0.570	1.063	1.898
3 Mean		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.000	0.008	0.005	0.010	0.000	0.000
Huber-1		Identified	2.000	3.000	4.000	0.000	0.000	0.000
		Swamped	5.640	2.303	0.525	0.058	0.000	0.000
Huber-2		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.025	0.000	0.000	0.000	0.000	0.000
Potential (mean)		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.005	0.013	0.010	0.015	0.000	0.000
Potential (med.)		Identified	1.970	2.790	3.470	3.525	5.100	5.050
		Swamped	0.450	0.363	0.325	0.268	0.155	0.103
Generalized Potential	c=3	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.150	0.145	0.150	0.173	0.138	0.120
	c=5	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.013	0.013	0.000	0.008	0.000	0.000

Table 5.c.3: Simulation results in presence of multiple (10%) equally high leverage point (4)

Measures		Identification Status	Sample size, *n*					
			20	30	40	50	100	200
2 Mean		Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.268	0.443	0.505	0.605	0.953	1.773
3 Mean		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.003	0.005	0.003	0.005	0.000	0.003
Huber-1		Identified	2.000	3.000	4.000	0.000	0.000	0.000
		Swamped	5.613	2.250	0.505	0.0400	0.000	0.000
Huber-2		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.035	0.000	0.000	0.000	0.000	0.000
Potential (mean)		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.003	0.007	0.003	0.005	0.003	0.003
Potential (med.)		Identified	1.975	2.850	3.490	3.862	5.150	5.400
		Swamped	0.433	0.395	0.323	0.265	0.130	0.070
Generalized Potential	c=3	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.163	0.170	0.170	0.175	0.130	0.128
	c=5	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.005	0.018	0.010	0.008	0.000	0.000

Table 5.c.4: Simulation results in presence of multiple (10%) equally high leverage point (5)

Measures		Identification Status	Sample size, *n*					
			20	30	40	50	100	200
2 Mean		Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.270	0.415	0.538	0.640	1.053	1.850
3 Mean		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.008	0.005	0.003	0.005	0.005	0.000
Huber-1		Identified	2.000	3.000	4.000	0.000	0.000	0.000
		Swamped	5.430	2.325	0.538	0.058	0.000	0.000
Huber-2		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.038	0.000	0.000	0.000	0.000	0.000
Potential (mean)		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.008	0.005	0.005	0.005	0.005	0.000
Potential (med.)		Identified	1.985	2.865	3.550	4.313	5.900	8.500
		Swamped	0.505	0.423	0.318	0.300	0.230	0.070
Generalized Potential	c=3	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.190	0.165	0.145	0.203	0.153	0.2113
	c=5	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.015	0.010	0.010	0.005	0.003	0.000

Table 5.c.5: Simulation results in presence of multiple (10%) equally high leverage point (8)

Measures		Identification Status	Sample size, *n*					
			20	30	40	50	100	200
2 Mean		Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.275	0.403	0.528	0.535	1.005	1.7625
3 Mean		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.013	0.000	0.003	0.003	0.000	0.000
Huber-1		Identified	2.000	3.000	4.000	0.000	0.000	0.000
		Swamped	5.465	2.203	0.528	0.045	0.000	0.000
Huber-2		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.045	0.000	0.000	0.000	0.000	0.000
Potential (mean)		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.013	0.003	0.005	0.003	0.000	0.000
Potential (med.)		Identified	1.995	2.865	3.620	4.288	6.200	8.200
		Swamped	0.445	0.393	0.303	0.240	0.110	0.063
Generalized Potential	c=3	Identified	2.000	3.000	4.000	10.000	10.000	20.000
		Swamped	0.180	0.168	0.165	0.123	0.123	0.140
	c=5	Identified	2.000	3.000	4.000	10.000	10.000	20.000
		Swamped	0.020	0.010	0.005	0.000	0.000	0.000

Table 5.c.6: Simulation results in presence of multiple (10%) equally high leverage point (10)

Measures		Identification Status	Sample size, *n*					
			20	30	40	50	100	200
2 Mean		Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.230	0.418	0.518	0.605	1.015	1.840
3 Mean		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.000	0.005	0.003	0.000	0.000	0.000
Huber-1		Identified	2.000	3.000	4.000	0.000	0.000	0.000
		Swamped	5.650	2.273	0.518	0.068	0.000	0.000
Huber-2		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.013	0.000	0.000	0.000	0.000	0.000
Potential (mean)		Identified	0.000	0.000	0.000	0.000	0.000	0.000
		Swamped	0.000	0.008	0.003	0.000	0.000	0.000
Potential (med.)		Identified	1.995	2.873	3.650	4.350	6.450	8.299
		Swamped	0.373	0.373	0.370	0.253	0.138	0.077
Generalized Potential	c=3	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.123	0.158	0.190	0.175	0.115	0.112
	c=5	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.018	0.013	0.003	0.008	0.000	0.000

We observe from Tables 5.c.1 to 5.c.6 that the performance of 3 Mean, Huber-2 and potential (mean) is very poor. All the methods could not detect a single high leverage point for different sample sizes and different weight of leverage points. They have a negligible swamping rate.

Huber-1 method is performed well for small samples and for each set of leverage points, For example in Table 5.a.1, this method identified on average 2, 3, 4 high leverage points for sample sizes 20, 30, and 40 where the magnitude of leverage points is 2. We also observed from this Table that the swamping rate of this method is high. For sample sizes 20, 30 and 40 it swamps 5.71, 2.425 and 0.563 observations respectively. That is the swamping rate decreases with the increases of sample sizes. But even for the moderate sample size like 50 it is not satisfactory. It could not identify even a single high leverage points for rest of the sample sizes. Similar results are found from the all other Tables.

Potential (med) is performed comparatively well for small samples, but its performance tends to become worse with the increase in sample sizes.

The performance of Twice the mean rule is satisfactory for all samples and different values of high leverage points. It correctly identifies the high leverage points. It has a considerable swamping rate.

From our simulation experiment, we observed that Generalized Potential performed best. It was able to identify all high leverage points correctly and produced very low swamping rate.

5.3 Identification of multiple (10%) unequally high leverage points

In this section we shall investigate how successfully the different measures of leverage identify the multiple high leverage points when in fact 10% of the observations are high leverage points and have different weights. The first 90% observations of the three regressor data set for sample sizes n = 20, 30, 40, 50, 100 and 200 are generated as Uniform (0,1) and the last 10% observations are constructed as high leverage points. To generate the high leverage values of unequal weights the values for each of the three regressors corresponding to the first high leverage point are kept fixed at 2 and those of the successive values have increments of 2. Ten thousand simulations are run for each of five set of measures and for each of six sample sizes and the results based on their averages are presented in Table 5.d.

Table 5.d: Simulation results in presence of multiple (10%) unequally high leverage points

Measures		Identification Status	Sample size, n					
			20	30	40	50	100	200
2 Mean		Identified	1.000	2.000	2.000	2.000	5.000	10.000
		Swamped	0.283	0.410	0.425	0.593	0.963	1.638
3 Mean		Identified	1.000	1.000	1.000	2.000	4.000	7.000
		Swamped	0.005	0.003	0.008	0.000	0.000	0.000
Huber-1		Identified	1.000	2.000	2.000	2.000	2.000	0.000
		Swamped	5.433	2.268	0.425	0.025	0.000	0.000
Huber-2		Identified	1.000	1.000	1.000	0.000	0.000	0.000
		Swamped	0.035	0.000	0.000	0.000	0.000	0.000
Potential (mean)		Identified	1.000	1.000	1.000	1.000	3.000	5.600
		Swamped	0.000	0.000	0.000	0.000	0.000	0.000
Potential (med.)		Identified	1.000	1.640	2.000	2.100	4.683	8.918
		Swamped	0.618	0.635	0.375	0.473	0.370	0.238
Generalized Potential	c=3	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.1900	0.178	0.170	0.145	0.150	0.145
	c=5	Identified	2.000	3.000	4.000	5.000	10.000	20.000
		Swamped	0.018	0.013	0.005	0.000	0.000	0.000

5.3.1 Simulation Result Discussion

From the table 5.c it is interesting to note that all the detection techniques accept Generalised Potential break down completely here. So far successful 2-mean rule breaks down here. For sample sizes 20, 30, 40, 0, 100 and 200, it identified 1, 2, 2, 2, 5 and 10 high leverage points respectively. Almost similar results are found for the other methods. The performance of Generalised Potential method is outstanding. For all the samples this method is correctly identified each high leverage points.

5.4 Graphical Display for Locating Multiple High Leverage Points

In this section we would like to introduce a new graphical display for locating multiple high leverage points together with outliers and influential observations. A good number of diagnostic plots are now available in the statistical literature [see Atkinson (1985), Gray (1986), Hadi (1992), Ghosh (1996), Tsai *et al.* (1998)] for various purposes. In this section we consider a class of plots that uses residuals and leverages of any observation simultaneously to assess the influence of them on the fit. Although the locating and testing of a single unusual case has been largely resolved by the use of single case deletion methods, these methods may be ineffective when a group of unusual cases are responsible for the poor fitting of the model and hence methods based on group deletions are required. Similar remarks may apply to the diagnostic plots, which involve residuals and leverages. We introduce two types of diagnostic plots, the first of which is based on the least squares residuals and leverage values, and the second plot is based on single case deleted residuals and leverages.

5.4.1 Leverage-Residual (L-R) Plot

Gray (1986) proposed the Leverage-Residual (L-R) plot, which is a simple graphical display of the leverage and residual values for each case in a regression data set. In this plot, the leverage value w_{ii} for each observation i, is plotted against the square of a normalised form of its corresponding residual $\hat{\epsilon}_i^{\,2} / \sum_{i=1}^{n} \hat{\epsilon}_i^{\,2}$. The bulk of the cases will be associated with low leverage and small residuals so that they cluster near (0, 0). The unusual cases will have either high leverages or large residual components and will tend to be separated from the bulk of the cases. High leverage cases will be located in the upper area of the plot and observations with large residuals will be located in the area to the right.

5.4.2 Potential-Residual (P-R) plot

Hadi (1992) pointed out that in the L-R plot the high leverage cases do not get a proper emphasis in comparison with the cases having large residuals. He proposed an alternative plot, which he named as the Potential-Residual (P-R) plot, where potentials are used as alternatives to the leverages. In a P-R plot, the potential value $p_{ii} = w_{ii}/(1-w_{ii})$ for each observation i, is plotted against a normalised residual component

$$\frac{k\,\hat{\epsilon}_i^{\,2}}{(1-w_{ii})\left(\sum \hat{\epsilon}_i^{\,2} - \hat{\epsilon}_i^{\,2}\right)}.$$

Both the L-R plot and the P-R plot could be useful in assessing influences of observations in a single case diagnostic study. The P-R plot gives more emphasis to high leverage cases than the L-R plot. But in a masking and/or swamping

situation both of them could produce misleading plots and therefore diagnostic plots based on group deletions are wanted.

5.4.3 Generalized Potentials-Deletion Residuals (GP-DR) Plot

It is now evident that the diagnostic plots discussed in the previous section could only be informative when the right choice of leverage and residual components is made. If any group deletion method is able to produce a better set of residuals which are free of masking and swamping effects, one might expect an appropriate graphical display when these are used in a plot with their corresponding leverage components free from the same effects like the generalized potentials.

Here we propose a simple graphical display of group deleted leverages and residuals. We would use generalised potentials as leverages and the robust RLS residuals as deletion residuals in this plot and call it 'generalised potentials - deletion residuals (GP-DR)' plot. Since the high leverage points need not to be outliers and outliers may not be points of high leverage we would expect different deletion sets D for the computation of these two quantities. The main advantage of the GP-DR plot is that it is suitable for the data where masking and/or swamping makes single case diagnostic plots misleading. It is also interesting to note that this plot, unlike the L-R and P-R plots retains the signs of residuals, which we believe is very important when their interpretation is considered. Another difference between the GP-DR plot and the other plots is that we do not propose the normalisation of residuals or leverages. It is quite possible to suitably normalise generalized potentials and deletion residuals so that they could be measured on a similar scale, but for plots it is not crucial, as they are only scale factors. Since the bulk of the cases will be associated with low leverage and small residuals so most

of the pairs $\left(\hat{\in}_i^{(-D)},\ p_{ii}^{*}\right)$ will cluster near (0, 0). The unusual cases will have either high leverages or large residual components and will tend to be separated from the bulk of the cases. High leverage cases will be located in the upper area of the plot and observations with large residuals will be located either in the area to the left or to the right depending on the sign of them.

Examples

Here we consider several well-known data sets, which are frequently referred to in the study of the identification of influential observations, high leverage points and outliers. It should be noted that a large body of data sets is now available in the literature for the same purpose, but in general we would prefer to consider data sets where we definitely know the real situation. Although such simulated data sets are artificial in nature, it may help us to investigate the performance of different methods more reliably. Otherwise there is always uncertainty [Cook and Hawkins (1990)] about which observations are actually unusual. For each of the examples considered in this section we would display three different plots; the L-R plot, the P-R plot and the GP-DR plot. For GP-DR plot we compute reweighted least squares residuals as deletion residuals (DR) by using PROGRESS program given by Rousseeuw and Leroy (1987) while the computation of generalized potentials (GP) are done with a simple program written in MINITAB.

Examlpe-1

First we would consider the well-known Hawkins, Bradu, and Kass (1984) data. It has been reported by many authors [Rousseeuw and Leroy (1987)] that all single case deletion methods not only fail to identify the true outliers but most of them also identify high leverage inliers as outliers. Imon (1996) pointed out that

commonly used leverage measures also fail to focus on all of the high leverage points. Table.d.1 presents leverages, potentials, generalized potentials, OLS residuals and RLS residuals for this data.

Table 5.e: OLS and RLS residuals for Hawkins-Bradu-Kass (1984) Data

Index	OLS	RLS	Index	OLS	RLS	Index	OLS	RLS
1	3.38	9.74	26	-0.48	-0.70	51	0.89	0.65
2	3.99	10.18	27	-1.37	-0.74	52	-1.16	-0.55
3	3.00	10.41	28	-0.24	0.41	53	-0.12	1.01
4	2.56	9.56	29	0.39	0.39	54	1.71	0.69
5	3.06	10.11	30	-1.27	-0.07	55	0.73	0.09
6	3.44	10.00	31	-0.27	-0.08	56	0.78	0.05
7	4.51	10.80	32	0.56	-0.34	57	0.62	0.74
8	3.84	10.38	33	-0.11	-0.59	58	0.28	-0.17
9	2.71	9.77	34	-0.68	-0.52	59	-0.74	-0.18
10	3.04	10.10	35	-0.40	0.29	60	-1.35	-0.63
11	-7.83	-0.06	36	-1.17	-0.86	61	-0.02	-0.13
12	-9.37	-0.20	37	-0.23	-0.41	62	0.69	0.72
13	-6.12	0.62	38	1.25	0.92	63	0.65	-0.31
14	-3.80	-0.21	39	-1.27	-0.63	64	-0.89	-0.52
15	-0.66	-0.50	40	-0.02	-0.38	65	-0.29	0.73
16	0.87	0.46	41	-1.10	-0.11	66	-0.26	-0.77
17	0.65	-0.07	42	-1.08	-0.44	67	-0.49	-0.67
18	-0.39	0.03	43	1.72	0.69	68	0.54	0.83
19	0.65	0.18	44	-1.80	-0.52	69	0.19	0.21
20	0.34	0.31	45	-0.76	-0.45	70	0.47	0.86
21	0.67	0.88	46	-1.43	-0.23	71	0.01	0.22
22	0.93	0.42	47	-1.50	-0.93	72	0.14	-0.07
23	-0.43	-0.83	48	-0.87	0.12	73	0.44	0.60
24	1.35	0.71	49	0.65	0.83	74	-0.39	-0.72
25	-0.30	-0.18	50	-0.69	-0.35	75	-0.35	0.47

It should be noted that to compute GP the first 14 observations of this data set were omitted while the first 10 observations were deleted to compute RLS residuals.

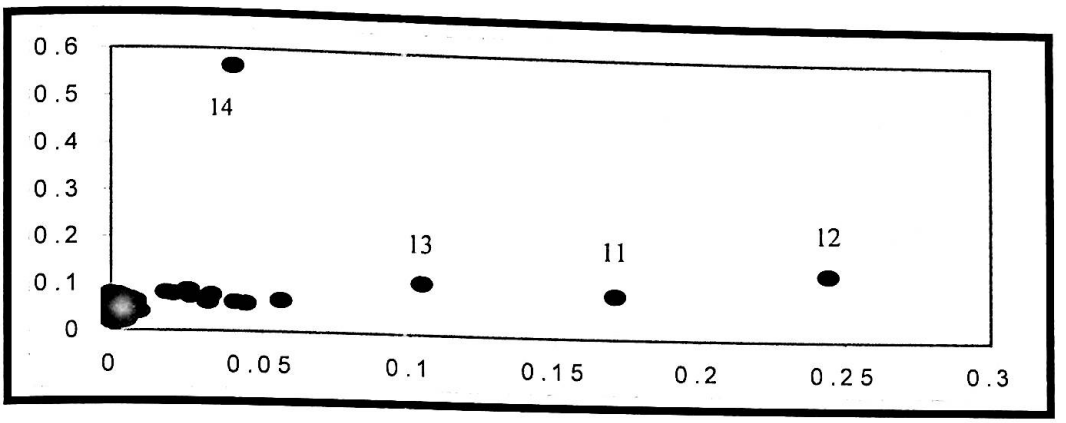


Figure 5.3.a. L-R Plot for Hawkins *et al.* (1984) data

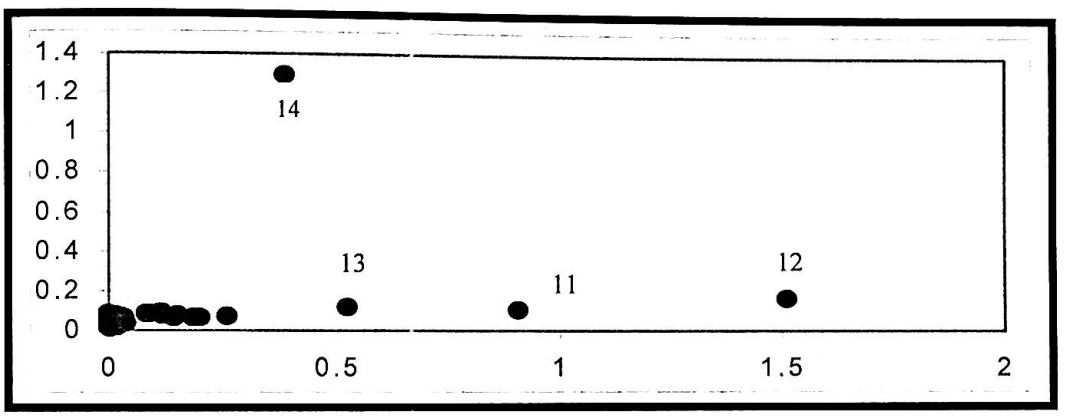


Figure 5.3.b. P-R Plot for Hawkins *et al.* (1984) data

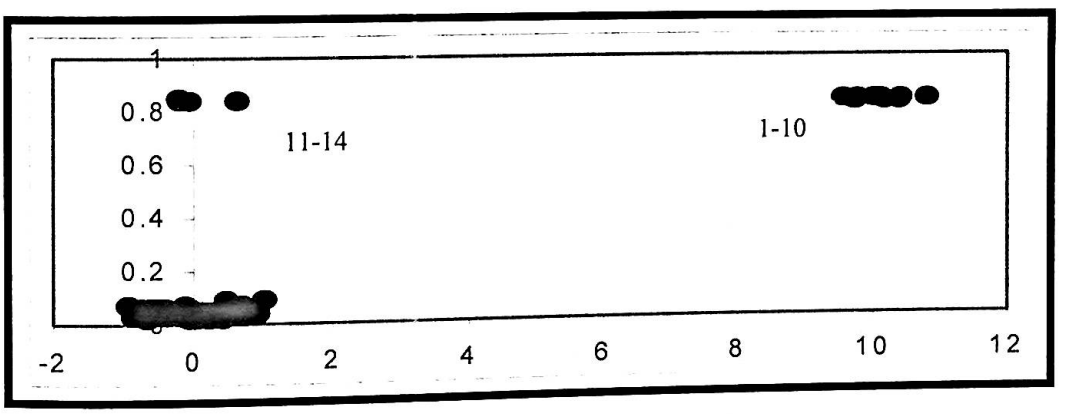


Figure 5.3.c. GP-DR Plot for Hawkins *et al.* (1984) data

Figures 5.3.a to 5.3.c show different diagnostic plots for the Hawkins *et al.* (1984) data. Both the L-R plot and the P-R plot fail to identify true outliers, which cluster near (0,0) with other 61 clean observations. Only one of the 14 high leverage points, *i.e.* the observation no. 14 is identified as the point of high leverage and unfortunately 3 high leverage inliers (cases 11-13) are identified as outliers. However, the newly proposed GP-DR plot becomes very successful in locating three groups of observations, which are clearly separated from one another. All of the clean observations cluster around (0,0), 10 high leverage outliers (with positive signs) are located in the top right corner of the plot and 4 high leverage inliers are located in the top of the center of this plot.

Example-2

Our next example is Peña and Yohai (1995) artificial data set-B. The main feature of this single predictor data set is that the two outlying observations (cases 9 and 10), which are also the points of equally high leverage, correspond to the true disturbances which are equal in magnitude but have opposite signs. The RLS and the GP are computed for this data set after the omission of these two high leverage outliers. The leverages, potentials, generalised potentials, OLS residuals and RLS residuals for Peña and Yohai (1995) data set-B are presented in Table 5.f.

Table 5.f: Leverages and residuals for Peña and Yohai (1995) artificial data set-B.

Case	w_{ii}	p_{ii}	p^*_{ii}	OLS	RLS
1	0.2894	0.4072	0.4167	0.021	0.025
2	0.2212	0.2840	0.2738	-0.081	-0.079
3	0.1682	0.2022	0.1786	-0.083	-0.082
4	0.1303	0.1498	0.1310	0.115	0.114
5	0.1076	0.1205	0.1310	0.212	0.211
6	0.1000	0.1111	0.1786	-0.090	-0.093
7	0.1076	0.1205	0.2738	-0.192	-0.196
8	0.1303	0.1498	0.4167	0.105	0.100
9	0.3727	0.5942	1.4640	5.996	5.986
10	0.3727	0.5942	1.4640	-6.004	-6.014

It is interesting to observe from this table-5.e. that the RLS and the OLS residuals for this data are almost identical. Because of balancing effect, these two outliers do not cause any damage to the fitting of the model that is why they are not jointly influential. Andrews and Pregibon (1978) have termed such cases as outliers that do not matter.

Different diagnostic plots for this data set are given in figures 5.4.a to 5.4.c.

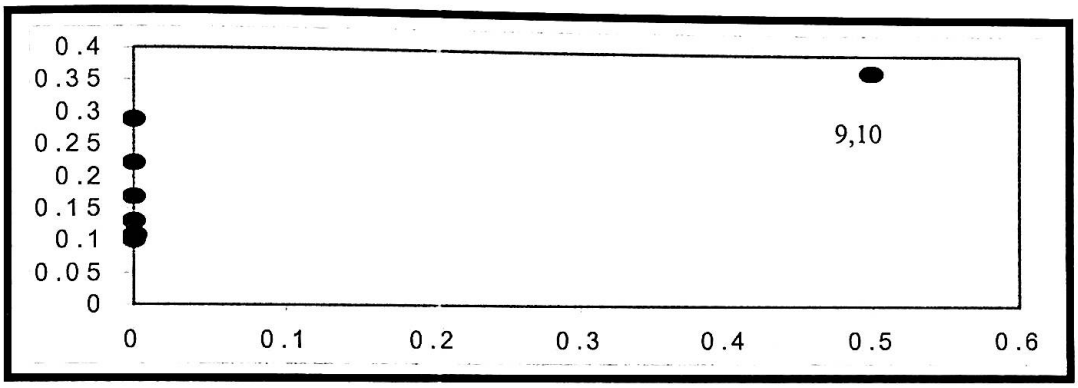


Figure- 5.4.a. L-R Plot for Peña-Yohai (1995) data set-B

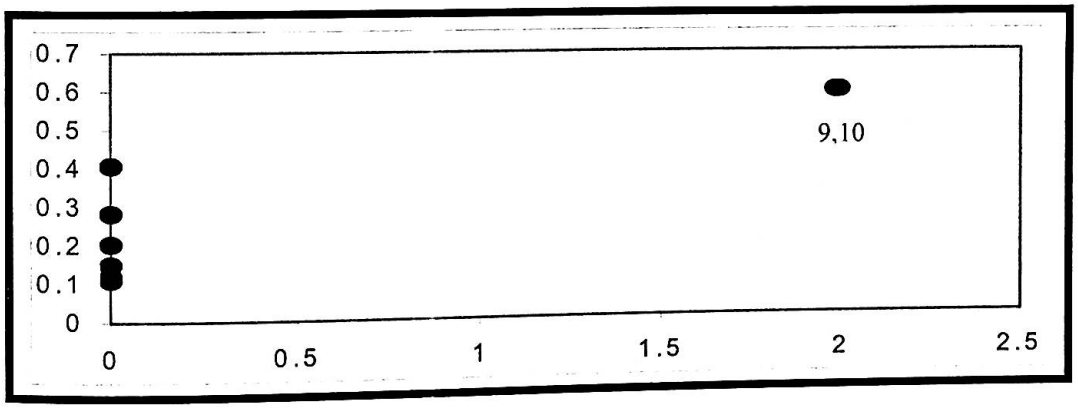


Figure 5.4.b. P-R Plot for Peña-Yohai (1995) data set-B

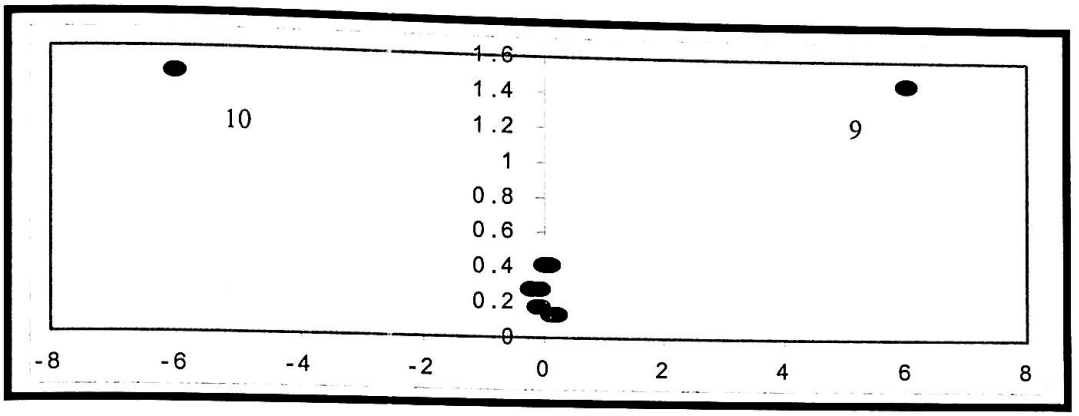
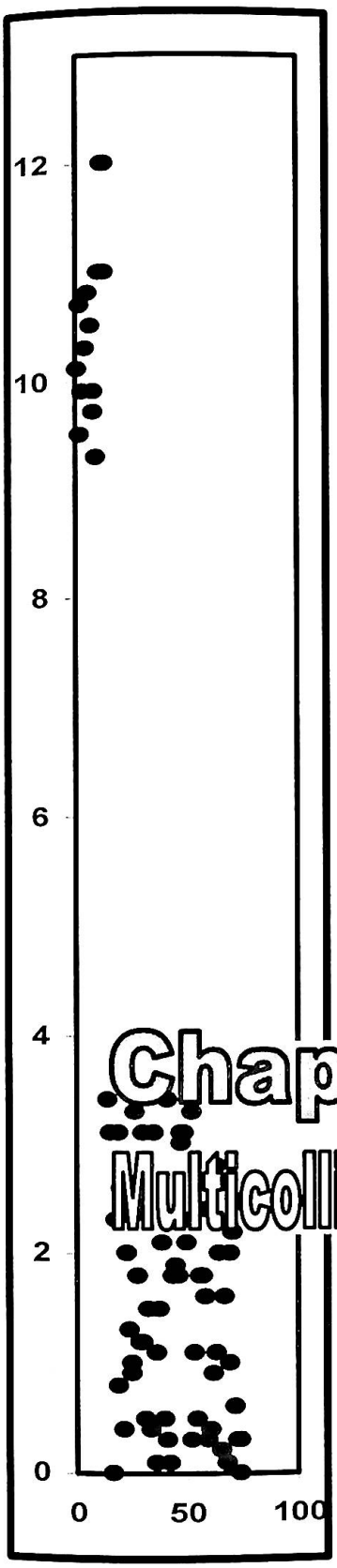


Figure 5.4.c. GP-DR Plot for Peña-Yohai (1995) data set-B

Both the L-R plot and the P-R plot can locate the two outliers successfully but the two high leverage points are not much focused on these plots. The GP-DR plot clearly focuses on the high leverage and the outlying behavior of cases 9 and 10, but more crucially they exhibit the balancing effect of the two outliers. Observation 9 appears in the top right corner of this plot while observation 10 is located in the top left corner of the plot indicating that those are high leverage outliers but because of their balancing effects they are outliers that do not matter. Neither the L-R plot nor the P-R plot can tell us that the two outliers are actually not jointly influential since both of them are located at the right corner of these plots. This example emphasises our concern that diagnostic plots should retain the signs of the residuals for the better interpretation of the results.



Chapter Six

Multicollinearity and High Leverage points

Chapter Six

Multicollinearity and High Leverage Points

Multicollinearity is considered as one of the most serious consequences when the assumptions regarding the OLS technique is violated. As we already mentioned that linear independence among the regressors is one of the fundamental assumptions of the OLS and violation of this assumption has a drastic consequence on the subsequent analysis. We suspect that the presence of high leverage point is responsible for causing multicollinearity. At first we discuss in a brief what is multicollinearity, the sources, consequences, detection techniques, and methods of dealing with multicollinearity. We observe how a single high leverage point causes multicollinearity. We also investigate whether the existing methods can successfully detect high leverage points or not and, the behavior of multicollinearity when high leverage points thus identified are omitted from the regression model. We also extend this experiment to the case when a group of high leverage points is present.

6.1 Concept of Multicollinearity

In a linear regression model, if there is no relationship between the regressors, they are said to be orthogonal. When the regressors are orthogonal, inferences such as **(a)** identifying the relative effects of the regressors, **(b)** prediction or estimation and **(c)** selection of an appropriate set of variables for the model, can be made relatively easily. Unfortunately in most applications of regression, the regressors are not orthogonal. Sometimes the lack of orthogonality is not serious. However, in some situations the regressors are nearly perfectly linearly related, and in such cases the inferences based on erroneous. When there are near linear dependencies between the regressors, the problem of multicolinearity is said to exit.

6.2 The Nature of multicolinearity

The term multicolinearity is due to Ragnar Frisch (1934). Originally it meant the existence of a "*perfect*" or exact linear relationship among some or all explanatory variables of a regression model.

For the k-variable regression involving explanatory variable $X_1, X_2, \cdots\cdots X_K$ (where $X_1 = 1$ for all observations to allow for the intercept term), an exact linear relationship is said to exit if the following condition is satisfied:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots\cdots + \lambda_K X_K = 0 \qquad (6.1)$$

where $\lambda_1, \lambda_2 \cdots\cdots \lambda_K$ are constant such that not all of them are zero simultaneously.

Today, however the term multicolinearity is used in a broader sense to include the case of perfect multicolinearity, as shown by (6.1), as well as the case where the X variables are intercorrelated but not perfectly so, as follows:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots\cdots + \lambda_K X_K + V_i = 0 \qquad 6.2)$$

Where V_i is a stochastic error term.

To see the difference between perfect and less than perfect multicolinearity, assume, for example, that $\lambda_2 \neq 0$, then equation (6.1) can be written as

$$X_{2i} = \frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots\cdots - \frac{\lambda_K}{\lambda_2} X_{Ki} \qquad (6.3)$$

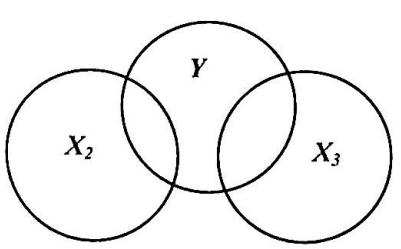
Which shows how X_2 is exactly linearly related to other variables or how it can be derived from a linear combination of other X variables. In this situation, the co-efficient of correlation between the variable X_2 and the linear combination on the right side of (6.3) is bound to be unity.

Similarly, if $\lambda_2 \neq 0$, equation (6.2) can be written as

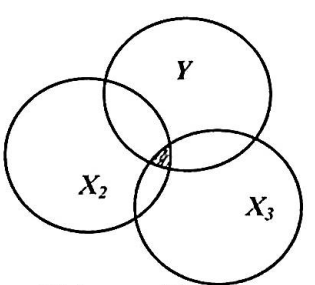
$$X_{2i} = \frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots\cdots - \frac{\lambda_K}{\lambda_2} X_{Ki} - \frac{1}{\lambda_2} V_i \qquad (6.4)$$

Which shows that X_2 is not an exact linear combinations of other X's because it is also determined by the stochastic error term V_i.

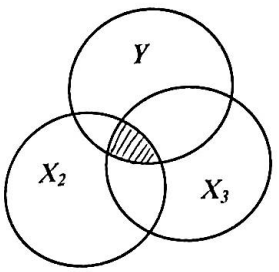
Kennedy (1981) suggested the Ballentine views of multicolinearity are as follows:



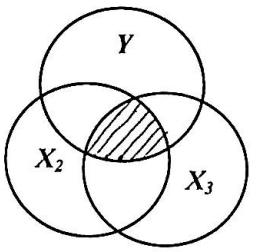
(a) No collinearity



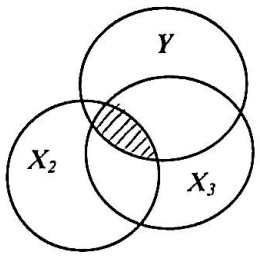
(b) Low collinearity



(c) Moderate collinearity



(d) High collinearity



(e) Very high collinearity

6.3 Sources of Multicolinearity

There are several sources of multicolinearity. As Montgomery and Peck (1992) note, multicollinearity may be due to the following factors:

(a) The data collection method employed, for example, sampling over a limited range of the values taken by the regressors in the population.

(b) Constraints on the model or in the population being sampled. For example, in the regression of electricity consumption on in come (X_2) and house size (X_3) there is a physical constraint in the population in that families with higher incomes generally have larger homes than families with lower incomes.

(c) Model specification, for example, adding polynomial terms to regression model, especially when the range of the X variable is small.

(d) An over defined model. This happens when the model has more explanatory variables than the number of observation. This could happen in the medical research where there may be a small number of patients about whom information is collected on a large number of variables.

6.4 Effects or consequences of Multicolinearity

The presence of near or high multicolinearity can occur the following problems:

(a) Although BLUE, the OLS estimators have large variances and covariances, making precise estimation difficult.

(b) Because of the consequence **(a)**, the confidence intervals tend to be much wider, leading to the acceptance of the 'zero null hypothesis' (i.e. the true population coefficient is zero) more readily.

(c) Also because of consequence of **(a)**, the t ratio of one or two coefficients tends to be statistically insignificant.

(d) Although the t ratio of one or more coefficients is statistically insignificant, R^2, the overall measure of goodness of fit, can be vary high.

(e) The OLS estimators and their standard errors can be sensitive to small changes in the data.

(f) The OLS estimators those are too large in absolute value.

6.5 Detection Techniques of Multicollinearity

Several techniques have been proposed for detecting multicollinearity. Some of them are,

(a) Examination of the correlation matrix.

(b) Variance inflation factors (VIF)

(c) Tolerance

(d) Condition number

(e) Eigen value decomposition

(a) Examination of the correlation matrix

A very simple suggested multicollinearity detection technique is that if the pair-wise correlation coefficient between two regressiors is high, say, in excess of 0.8, the multicollinearity is a serious problem.

(b) Variance Inflation Factor (VIF)

The diagonal elements of the $C = (X^T X)^{-1}$ matrix are very useful in detecting multicolinearity. We see that C_{jj}, the j-th diagonal elements of C can be written as $C_{jj} = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of determination obtained when x_j is regressed on the remaining (k-1) regressors. Marquardt (1970) has called $C_{jj} = (1-R_j^2)^{-1}$, the "*Variance Inflation Factor (VIF).*"

The VIF for each term in the model measures the combined effect of the dependencies among the regressors on the variance of that term. One or more large VIF's exceeds 5 or 10; it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

(c) Tolerance

In recent times tolerance values are suggested [see Sen and Srivastava (1990)] to use as a measure of multicollinearity. Tolarences are in fact the inverse of the VIF values, i.e. $TOL_j = (1-R_j^2)$. Higher R_j^2 values lead to multicollinearity that is why lower values of tolerance are undesirable. Tolerance value less then 0.1 indicates a strong multicollinearity, whereas values between 0.1 and 0.2 indicates moderate multicollinearity.

(d) Condition Number

The characteristic roots or eigen values of (X^TX), say $\lambda_1, \lambda_2 \cdots\cdots \lambda_p$ can be used to measure the exact of multicollinearity in the data. Some analysis perform to examine the condition number of defined as

$$\eta = \frac{\lambda_{max}}{\lambda_{min}}$$

This is just a number of the spread in the eigen values spectrum of (X^TX). Generally if the condition number is less than 100, there is no serious problem with multicollinearity, condition numbers (η) between 100 and 1000 imply moderate to strong multicollinearity and if η exceeds 1000 serious multicollinearity is indicated.

(e) Eigen or Singular Value Decomposition

Eigen system analysis can be used to identify the nature of the near linear dependencies in the data. The (X^TX) matrix may be decomposed as

$$X^TX = T\Lambda T^T$$

Where Λ is a $k \times k$ diagonal matrix whose main diagonal elements are the eigen values $\lambda_j \ (j = 1, 2, \cdots k)$ of (X^TX) and T is a $(k \times k)$ orthogonal matrix whose columns are the eigen vectors of (X^TX). Let the columns of T denoted by $t_1, t_2, \cdots\cdots t_k$. If the eigen value λ_j is close to zero, indicating a near linear dependency in the data, the elements of the associated eigen vector t_j describe the nature of this linear dependency. Belsley, Kuh and Welsch (1980) purpose a similar approach for diagnosing multicollinearity. Then X matrix of order $n \times k$ may be decomposed as

$$X = UDT^T$$

Where U is $(n \times k)$, T is $(k \times k)$, $UU^T = I$ and D is a $(k \times k)$ diagonal matrix with non-negative diagonal elements μ_j, $j = 1, 2, \cdots k$. The μ_j are called the singular values of X. The singular-value decomposition is closely related to the concepts of eigen values and eigen vectors, since $X^T X = \left(UDT^T\right)^T UDT^T = TD^T DT^T = T\Lambda T^T$, so that the squares of the singular values of X are the eigen values of $(X^T X)$. T is the matrix of eigen vectors of $(X^T X)$ defined earlier, and U is a matrix whose columns are the eigen vectors associated with the k nonzero eigen values of $(X^T X)$.

The covariance matrix of $\hat{\beta}$ is

$$V\left(\hat{\beta}\right) = \sigma^2\left(X^T X\right)^{-1} = \sigma^2 T\Lambda^{-1}T^T$$

and the variance of the j-th regression coefficient is the j-th diagonal element of this matrix, or

$$V\left(\hat{\beta}_j\right) = \sigma^2 \sum_{i=1}^{k} \frac{t_{ji}^2}{\mu_j^2} = \sigma^2 \sum_{i=1}^{k} \frac{t_{ji}^2}{\lambda_i}$$

Note also that apart from σ^2, the j-th diagonal element of $T\Lambda^{-1}T^T$ is the j-th variance inflation factor, so

$$VIF_j = \sum_{i=1}^{k} \frac{t_{ji}^2}{\mu_j^2} = \sum_{i=1}^{k} \frac{t_{ji}^2}{\lambda_j}$$

Clearly, one or more small singular values (or small eigen values) can dramatically inflate the variance of $\hat{\beta}_j$. Belsley, Kuh and Welsch (1980) suggest using variance decomposition proportions, for example

$$\pi_{ij} = \frac{\left(t_{ji}^2 / \mu_j^2\right)}{VIF_j}, \quad i, j = 1, 2, \cdots\cdots k$$

as measures of multicollinearity. If a high proportion of the variance for two or more regression coefficients is associated with one small singular value, multicolinearity is indicated. For example if π_{32} and π_{34} are large, the third singular value is associated with a multicollinearity. Variance decomposition proportions greater than 0.5 are recommended guidelines.

6.6 Methods for dealing with multicolinearity

Several techniques have been proposed for dealing with the problems caused by multicolinearity. The general approaches include,

(a) Collecting additional data.
(b) Model respecification
(c) Ridge regression
(d) Principle components regression
(e) Transformation of variables.
(f) Reducing Collinearity in polynomial regression.
(g) Combining cross sectional and time series data
(h) A priori information.

(a) Collecting additional data

Collecting additional data has been suggested as the best method of combating multicollinearity [see Farrar and Glauber (1967) and Silvey (1969)]. The additional data should be collected in a manner designed to break up the multicollinearity in the existing data.

(b) Model respecification

Multicollinearity is often caused by the choice of model, such as when two highly correlated regressors are used in the regression equation. In these situations some re-specification of the regression equation may lessen the impact of multicollinearity

(c) Ridge regression

When the method of least squares is applied to nonorthogonal data, very poor estimates of the regression coefficients are usually obtained. The problem with the method of least squares is the requirement that $\hat{\beta}$ be an unbiased estimator of β. The Gauss-Markoff property assures us that the least squares estimator has minimum variance in the class of unbiased linear estimators, but there is no guarantee that this variance will be small. One way to alleviate this problem is to drop the requirement that the estimator of β be unbiased.

A number of procedures have been developed for obtaining biased estimators of regression coefficiants. One of these procedures is ridge regression, originally proposed by Hoerl and Kennard (1970). The ridge estimator is found by solving a slightly modified version of the normal equations. Specifically, we define the ridge estimator $\hat{\beta}_R$ as the solution to $\left(X^T X + pI\right)\hat{\beta}_R = X^T Y$

or $$\hat{\beta}_R = \left(X^T X + pI\right)^{-1} X^T Y$$

where $p \geq 0$ is a constant selected by the analyst.

Now it is shown that

$$\text{MSE}(\hat{\beta}_R) = \sigma^2 \sum_{j=1}^{k} \frac{\lambda_j}{(\lambda_j + p)^2} + p^2 \beta^T (X^T X + kI)^{-2} \beta \cdots\cdots \quad (6.12)$$

Where $\lambda_1\ \lambda_2 \cdots\cdots \lambda_k$ are the eigen values of $(X^T X)$. The first term on the right-hand side of (6.12) is the sum of the parameters in $\hat{\beta}_R$ and the second term is the square of the bias. If $p > 0$, note that the bias in $\hat{\beta}_R$ increases with p. However, the variance decreases as p increases.

(d) Principal Components Regression

Biased estimators of regression coefficients can also be obtained by using a procedure known as principal components regression. Consider the canonical form of the model,

$$y = Z\alpha + \varepsilon$$

Where $Z = XA, \quad \alpha = A^T \beta,$ and $A^T X^T XA = Z^T Z = \Lambda$

Recall that $\Lambda = \text{diag}(\lambda_1\ \lambda_2 \cdots\cdots \lambda_k)$ is a k×k diagonal matrix whose columns are the eigen vectors associated with $\lambda_1\ \lambda_2 \cdots\cdots \lambda_k$. The columns of Z, which define a new set of orthogonal regressors, such as $Z = [Z_1, Z_2, \cdots\cdots, Z_k]$ are referred to as principal components.

The least squares estimator of α is

$$\hat{\alpha} = (Z^T Z)^{-1} Z^T y = \Lambda^{-1} Z^T y$$

and the covariance matrix of $\hat{\alpha}$ is

$$V(\hat{\alpha}) = \sigma^2 \left(Z^T Z\right)^{-1} = \sigma^2 \Lambda^{-1}$$

Thus a small eigen value of ($X^T X$)means that the variance of the corresponding orthogonal regression coefficient will be large. Since

$$Z^T Z = \sum_{i=1}^{k} \sum_{j=1}^{k} Z_i Z_j^T = \Lambda$$

We often refer to the eigen value λ_j as the variance of the j-th principal component. If all the λ_j are equal to unity, the original regressors are orthogonal, while if an λ_j is exactly equal to zero this implies a perfect linear relationship between the original regressors. One or more of the λ_j near zero implies that multicollinearity is present.

The principal components regression approach combats multicollinearity by using less than the full set of principal components in the model. To obtain the principal components estimator, assume that the regressors are arranged in order of decreasing eigen values, $\lambda_1 \geq \lambda_2 \geq \cdots\cdots \geq \lambda_k > 0$. Suppose that the last s of the eigen values are approximately equal to zero. In principal components regression the principal components corresponding to near zero eigen values are removed from the analysis and least squares applied to the remaining components. That is,

$$\hat{\alpha}_{pc} = B\hat{\alpha}$$

where $b_1 = b_2 = \cdots = b_k$ and $b_{k-s-1} = b_{k-s-2} = \cdots = b_k = 0$.

Thus the principal components estimator is $\hat{\alpha}_{pc} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{k-s} \\ \cdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

or in terms of the standardized regressors

$$\hat{\beta}_{pc} = A\hat{\alpha}_{pc}$$

$$= \sum_{j=1}^{k-s} \lambda_j^{-1} a_j^T X^T y a_j$$

A simulation study by Gunst and Mason (1977) showed that principal components regression offers considerable improvement over least squares when the data are ill-conditioned.

Now the above remedial measures, the first four measures discussed in detail and the remaining four measures such as transformation of variables; reducing Collinearity in polynomial regression; a priori information and combining cross sectional and time series data discussed detail in Gujarati (1995, pp.340-44).

We anticipate that high leverage points may be another source of multicollinearity, but so far as we know no detection technique of multicollinearity has been developed on the high leverage issue. In the next section, we will try to address this issue and establish high leverage points as a source of multicollinearity.

6.7 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 to reduce the effect of multicollinearity. Here we present an example in favor of our proposition. We consider again data set which is presented in Table 5.a.1

Table 5.a.2 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 14-*th* 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 (13) of the previous section we observe that the first 14 observations are appearing as points of high leverages. The generalized potential values presented in Table 6.7.a are thus obtained from (12) with cases 1-14 deleted. This table also shows that the generalized 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.

Table 6.7.a: Multicollinearity diagnostics for Hawkins *et al.* data

Data	Correlation	Eigen Value	Condition Index	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. Lev.GP. (n=61)	r_{12}= 0.044	3.383	3.434	0.76	0.22	0.07
	r_{13}= 0107	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				

These results are presented in the above Table 6.7.a considers 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, condition indices 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 condition indices 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 do not help in the identification of high leverage points and

consequently we observe a little improvement in the results of multicollinearity. But the use of generalized potentials produces stunning results. When the high leverage points identified by this method are omitted from the analysis, we observe that the correlation values among X_1, X_2 and X_3 are very low. We also observe that neither of the condition indices 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 the data and none of the three variables is affected by it.

Now we present some 3D plots of explanatory variables that will show how generalized potentials contribute in handling the multicollinearity problem.

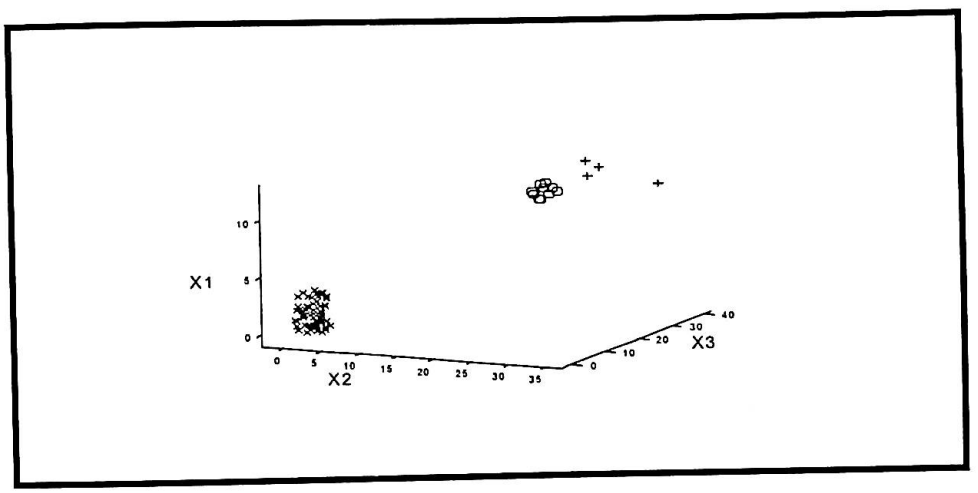


Figure 6.a. 3D plot of the original *X*'s of Hawkins *et al.* data

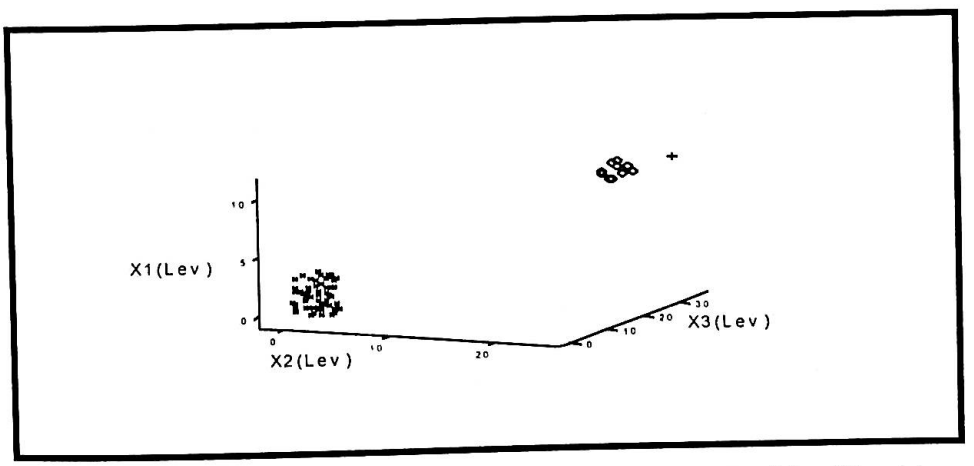


Figure 6.b. 3D plot of the *X*'s after deleting the cases by 2M method for Hawkins *et al.* data

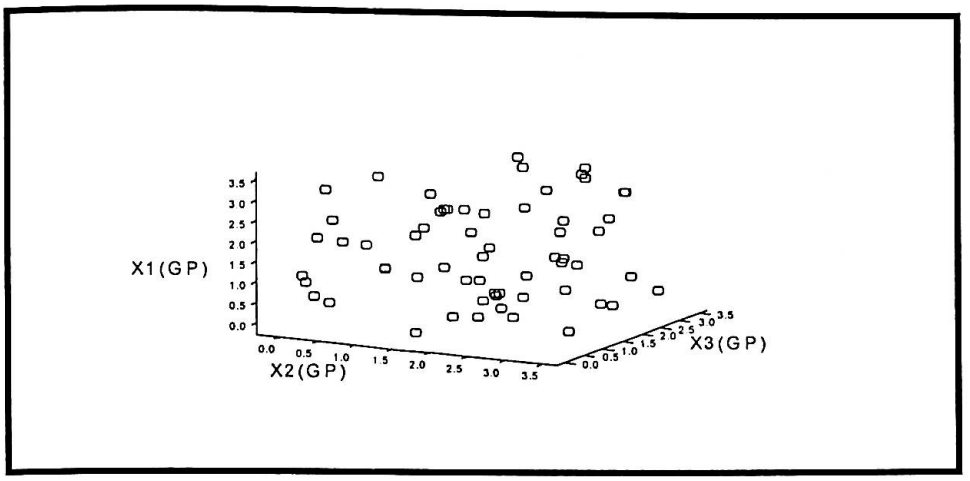


Figure 6.c. 3D plot of the X's after deleting the cases by GP method for Hawkins *et al.* data

Figure-6.a. presents a 3D plot of the X's with the original data. As we know that there are 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-6.b. where 3 out of 14 high leverage points are omitted. Figure 6.c. Presents 3D plot of the X's where high leverage points detected by generalized 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.

6.8 Simulation Results

Here we report a Monte Carlo simulation study, which is designed to investigate how high leverage points behave as a source of multicollinearity. At first we demonstrate how a single high leverage point causes multicollinearity. We also investigate whether the existing methods can successfully detect high leverage points or not and, the behaviour of multicollinearity when high leverage point thus

identified is omitted from the regression model. Then we extend this experiment to the cases where multiple high leverage points are present in the data. Likewise the examples considered earlier we generate three-predictor artificial data set for a single high leverage and multiple (10%) high leverage cases. 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 and 10) and pairwise correlation coefficient for these there sets of variable are computed. Throughout our simulation experiment we use seven detection techniques to identify high leverage points. Correlation coefficients r_{12}, r_{13} and r_{23} are computed together with result after omitting the observations identified by these seven detection techniques. Each of which is based on 10000 simulations.

6.8.1 Simulation Results for a Single High Leverage Point

In this subsection we report the simulation results where the X variables contain a single high leverage point have equal weight. The first (n-1) observations for each of the three explanatory variables are simulated as Uniform (0, 1). The n-th observation for each of the X's is kept fixed at a same high leverage value. Correlation coefficients of X's together with the results after excluding the suspect high leverage cases are presented in Tables 6.a.1-6.a.6 that are based on the average of 10,000 simulations.

Table 6.a.1: Simulation results in presence of single high leverage point = 2

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.59185	0.48318	0.41129	0.35741	0.21549	0.21365
	2 Mean	0.01471	0.00917	0.00838	0.01073	0.00817	0.18406
	3 Mean	0.00194	-0.00102	-0.00028	0.00299	0.00177	0.05641
	Huber-1	0.25331	0.06069	0.00838	0.00342	0.00177	-0.00163
	Huber-2	0.00348	-0.00138	-0.00046	0.00299	0.18942	0.18324
	Potential (mean)	0.00141	-0.00138	0.00287	0.00299	0.00177	-0.00163
	Potential (med)	0.03943	0.02295	0.02078	0.01954	0.00996	0.00505
	G.P.	0.00141	-0.00138	-0.00046	0.00299	0.00177	-0.00163
r_{23}	ACTUAL	0.58894	0.48377	0.41203	0.35640	0.21445	0.21209
	2 Mean	0.01321	0.01049	0.01178	0.00551	0.00507	0.18162
	3 Mean	-0.00006	-0.00049	0.00078	-0.00076	-0.00040	0.05268
	Huber-1	0.24691	0.06252	0.01178	0.00005	-0.00040	-0.00319
	Huber-2	0.00195	-0.00091	-0.00046	-0.00076	0.18774	0.18207
	Potential (mean)	-0.00028	-0.00091	0.00081	-0.00076	-0.00040	-0.00319
	Potential (med)	0.03635	0.02592	0.02269	0.01374	0.00655	0.00510
	G.P.	-0.00028	-0.00091	0.00081	-0.0076	-0.00040	-0.00319
r_{23}	ACTUAL	0.59462	0.48260	0.41209	0.35900	0.21407	0.21308
	2 Mean	0.02311	0.00678	0.01032	0.01262	0.00491	0.18819
	3 Mean	0.00846	-0.00138	0.00296	0.00459	-0.00055	0.05116
	Huber-1	0.25684	0.05852	0.01032	0.00498	-0.00055	-0.00216
	Huber-2	0.01054	-0.00120	0.00081	0.00459	0.18742	0.18302
	Potential (mean)	0.00818	-0.00120	0.00287	0.00459	-0.00055	-0.00216
	Potential (med)	0.04863	0.02180	0.02401	0.02218	0.00580	0.00456
	G.P.	0.00818	-0.00120	0.00287	0.00459	-0.00055	-0.00216

Table 6.a.2: Simulation results in presence of single high leverage point = 3

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.79995	0.72601	0.66172	0.60579	0.43282	0.27242
	2 Mean	0.02160	0.02250	0.01561	0.00663	0.00671	0.00141
	3 Mean	0.00952	0.01181	0.00709	-0.00198	0.00157	-0.00249
	Huber-1	0.25241	0.07115	0.01561	-0.00150	0.00157	-0.00249
	Huber-2	0.01205	0.01181	0.00709	-0.00198	0.00157	0.01169
	Potential (mean)	0.00903	0.01181	0.00709	-0.00198	0.00157	-0.00249
	Potential (med)	0.05053	0.03883	0.03044	0.01940	0.01447	0.00302
	G.P.	0.009903	0.01181	0.00709	-0.00198	0.00157	-0.00249
r_{23}	ACTUAL	0.79770	0.72339	0.66075	0.60540	0.43297	0.27339
	2 Mean	0.01116	0.01459	0.01139	0.00550	0.00776	0.00273
	3 Mean	-0.00264	0.00663	0.00453	-0.00174	0.00165	-0.00132
	Huber-1	0.24749	0.06308	0.01139	-0.00118	0.00165	-0.00132
	Huber-2	0.00016	0.00663	0.00453	-0.00174	0.00165	0.01275
	Potential (mean)	-0.00270	0.00663	0.00453	-0.00174	0.00165	-0.00132
	Potential (med)	0.03819	0.03143	0.02547	0.01950	0.01546	0.00507
	G.P.	-0.00270	0.00663	0.00453	-0.00174	0.00165	-0.00132
r_{23}	ACTUAL	0.79897	0.72314	0.65867	0.60858	0.43188	0.27813
	2 Mean	0.01704	0.01447	0.00706	0.01151	0.00374	0.00994
	3 Mean	0.00336	0.00374	-0.00339	0.00305	-0.00048	0.00528
	Huber-1	0.25536	0.06555	0.00706	0.00371	-0.00048	0.00528
	Huber-2	0.00469	0.00374	-0.00339	0.00305	-0.00048	0.01934
	Potential (mean)	0.00336	0.00374	-0.00339	0.00305	-0.00048	0.00528
	Potential (med)	0.04569	0.03068	0.02005	0.02426	0.01170	0.01183
	G.P.	0.00336	0.00374	-0.00339	0.00305	-0.00048	0.00528

Table 6.a.3: Simulation results in presence of single high leverage point = 4

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.88620	0.83492	0.79112	0.75096	0.59891	0.42495
	2 Mean	0.00866	0.00227	0.00627	0.00708	0.00542	0.00357
	3 Mean	-0.00454	-0.00837	-0.00099	0.00104	0.00235	-0.00018
	Huber-1	0.24717	0.05136	0.00627	0.00127	0.00235	-0.00018
	Huber-2	-0.00305	-0.00835	-0.00099	0.00104	0.00235	-0.00018
	Potential (mean)	-0.00459	-0.00835	-0.00099	0.00104	0.00235	-0.00018
	Potential (med)	0.03277	0.01665	0.01880	0.01703	0.01348	0.00869
	G.P.	-0.00459	-0.00835	-0.00099	0.00104	0.00235	-0.00018
r_{23}	ACTUAL	0.88697	0.83527	0.79067	0.75046	0.59960	0.42579
	2 Mean	0.02352	0.01188	0.00736	0.00660	0.00760	0.00360
	3 Mean	0.01217	0.00179	-0.00136	-0.00088	0.00266	0.00019
	Huber-1	0.26894	0.05815	0.00736	-0.00002	0.00266	0.00019
	Huber-2	0.01406	0.00165	-0.00136	-0.00088	0.00266	0.00019
	Potential (mean)	0.01140	0.00165	-0.00136	-0.00088	0.00266	0.00019
	Potential (med)	0.05507	0.02698	0.02019	0.01654	0.01601	0.00857
	G.P.	0.01140	0.00165	-0.00136	-0.00088	0.00266	0.00019
r_{23}	ACTUAL	0.88647	0.83766	0.79103	0.74915	0.59703	0.42532
	2 Mean	0.01600	0.02067	0.01132	-0.00077	0.00209	0.00311
	3 Mean	0.00810	0.01079	0.00323	-0.00654	-0.00218	0.00013
	Huber-1	0.24683	0.06811	0.01132	-0.00557	-0.00218	0.00013
	Huber-2	0.00358	0.01035	0.00323	-0.00654	-0.00218	0.00013
	Potential (mean)	0.00149	0.01035	0.00323	-0.00654	-0.00218	0.00013
	Potential (med)	0.04485	0.03767	0.02220	0.00849	0.00980	0.00893
	G.P.	0.00149	0.01035	0.00323	-0.00654	-0.00218	0.00013

Table 6.a.4: Simulation results in presence of single high leverage point = 5

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.92744	0.89201	0.86171	0.83336	0.70927	0.54907
	2 Mean	0.00035	0.00217	0.00897	0.00582	0.00036	0.00070
	3 Mean	-0.01292	-0.00833	0.00019	0.00062	-0.00375	-0.00242
	Huber-1	0.23364	0.05385	0.00897	0.00094	-0.00375	-0.00242
	Huber-2	-0.01208	-0.00842	-0.00018	0.00054	-0.00375	-0.00242
	Potential (mean)	-0.01275	-0.00842	-0.00018	0.00054	-0.00375	-0.00242
	Potential (med)	0.02184	0.01970	0.01913	0.01739	0.00733	0.00644
	G.P.	-0.01275	-0.00842	-0.00018	0.00054	-0.00375	-0.00242
r_{23}	ACTUAL	0.92943	0.89272	0.86194	0.83241	0.70997	0.55048
	2 Mean	0.03357	0.00694	0.00845	0.00563	0.00401	0.00404
	3 Mean	0.02035	-0.00090	-0.00030	-0.00242	-0.00124	0.00085
	Huber-1	0.27303	0.05334	0.00845	-0.00227	-0.00124	0.00085
	Huber-2	0.02201	-0.00111	-0.00021	-0.00237	-0.00124	0.00085
	Potential (mean)	0.02005	-0.00111	-0.00021	-0.00237	-0.00124	0.00085
	Potential (med)	0.06011	0.02208	0.01824	0.01599	0.00984	0.01009
	G.P.	0.02005	-0.00111	-0.00021	-0.00237	-0.00124	0.00085
r_{23}	ACTUAL	0.92782	0.89289	0.86176	0.83211	0.71116	0.54972
	2 Mean	0.01650	0.00667	0.01273	0.00441	0.00524	0.00272
	3 Mean	0.00290	-0.00419	0.00449	-0.00312	0.00105	-0.00044
	Huber-1	0.24838	0.05803	0.01273	-0.00255	0.00105	-0.00044
	Huber-2	0.00383	-0.00479	0.00437	-0.00320	0.00105	-0.00044
	Potential (mean)	0.00232	-0.00479	0.00437	-0.00320	0.00105	-0.00044
	Potential (med)	0.04217	0.02648	0.02914	0.01556	0.01123	0.00842
	G.P.	0.00232	-0.00479	0.00437	-0.00320	0.00105	-0.00044

Table 6.a.5: Simulation results in presence of single high leverage point = 8

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.97241	0.95903	0.94585	0.93262	0.87124	0.77264
	2 Mean	0.00083	0.01626	0.01597	0.01111	-0.00043	0.00504
	3 Mean	-0.01210	0.00705	0.00751	0.00484	-0.00467	0.00141
	Huber-1	0.24102	0.06866	0.01597	0.00546	-0.00467	0.00141
	Huber-2	-0.01025	0.00684	0.00751	0.00484	-0.00467	0.00141
	Potential (mean)	-0.01223	0.00684	0.00751	0.00484	-0.00467	0.00141
	Potential (med)	0.02178	0.03224	0.02614	0.02164	0.00385	0.00810
	G.P.	-0.01223	0.00684	0.00751	0.00484	-0.00467	0.00141
r_{23}	ACTUAL	0.97295	0.95891	0.94523	0.93199	0.87158	0.77231
	2 Mean	0.01816	0.01593	0.00541	0.00286	-0.00057	0.00256
	3 Mean	0.00690	0.00737	-0.00189	-0.00473	-0.00450	-0.00060
	Huber-1	0.24986	0.06495	0.00541	-0.00434	-0.00450	-0.00060
	Huber-2	0.00889	0.00731	-0.00189	-0.00473	-0.00450	-0.00060
	Potential (mean)	0.00659	0.00731	-0.00189	-0.00473	-0.00450	-0.00060
	Potential (med)	0.04219	0.03453	0.01562	0.01063	0.00324	0.00556
	G.P.	0.00659	0.00731	-0.00189	-0.00473	-0.00450	-0.00060
r_{23}	ACTUAL	0.97245	0.95890	0.94522	0.93237	0.87270	0.77215
	2 Mean	0.01161	0.00774	0.00454	0.00332	0.00776	0.00105
	3 Mean	-0.00292	-0.00066	-0.00417	-0.00273	0.00312	-0.00134
	Huber-1	0.24016	0.06207	0.00454	-0.00255	0.00312	-0.00134
	Huber-2	-0.00109	-0.00077	-0.00417	-0.00273	0.00312	-0.00134
	Potential (mean)	-0.00301	-0.00077	-0.00417	-0.00273	0.00312	-0.00134
	Potential (med)	0.03442	0.02500	0.01494	0.01296	0.01206	0.00408
	G.P.	-0.00301	-0.00077	-0.00417	-0.00273	0.00312	-0.00134

Table 6.a.6: Simulation results in presence of single high leverage point = 10

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.98300	0.97357	0.97357	0.96533	0.95709	0.91654
	2 Mean	0.02055	-0.00103	-0.00103	0.00981	0.00953	0.00345
	3 Mean	0.00746	-0.01152	-0.01152	0.00171	0.00283	-0.00011
	Huber-1	0.26028	0.04918	0.04918	0.00981	0.00334	-0.00014
	Huber-2	0.00964	-0.01169	-0.01169	0.00171	0.00283	-0.00014
	Potential (mean)	0.00688	-0.01169	-0.01169	0.00171	0.00283	-0.00014
	Potential (med)	0.04408	0.01178	0.01178	0.02230	0.01724	0.00658
	G.P.	0.00688	-0.01169	-0.01169	0.00171	0.00283	-0.00014
r_{23}	ACTUAL	0.98275	0.97410	0.97410	0.96533	0.95682	0.91637
	2 Mean	0.00395	0.02227	0.02227	0.01115	0.00809	0.00489
	3 Mean	-0.00714	0.01320	0.01320	0.00333	0.00205	0.00020
	Huber-1	0.24065	0.07378	0.07378	0.01115	0.00240	0.00016
	Huber-2	-0.00572	0.01308	0.01308	0.00333	0.00205	0.00016
	Potential (mean)	-0.00717	0.01308	0.01308	0.00333	0.00205	0.00016
	Potential (med)	0.02863	0.03839	0.03839	0.02380	0.01656	0.00899
	G.P.	-0.00717	0.01308	0.01308	0.00333	0.00205	0.00016
r_{23}	ACTUAL	0.98273	0.97412	0.97412	0.96554	0.95634	0.91704
	2 Mean	0.00935	0.01832	0.01832	0.01398	-0.00058	0.00997
	3 Mean	-0.00330	0.00730	0.00730	0.00503	-0.00803	0.00568
	Huber-1	0.24971	0.06760	0.06760	0.01398	-0.00736	0.00572
	Huber-2	-0.00286	0.00719	0.00719	0.00503	-0.00803	0.00572
	Potential (mean)	-0.00336	0.00719	0.00719	0.00503	-0.00803	0.00572
	Potential (med)	0.03337	0.03338	0.03338	0.02475	0.00762	0.01425
	G.P.	-0.00336	0.00719	0.00719	0.00503	-0.00803	0.00572

6.8.2 Simulation Results Discussion for a Single High Leverage Point

In Tables 6.a.1 to 6.a.6 it is observed that for every n, the presence of a single high leverage point 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 for example, in Table-6.a.1, the values of r_{12}, are 0.59185, 0.48318, 0.41129, 0.35741, 0.21549 and 0.21365, for n = 20, 30, 40, 50, 100 and 200 respectively. It is also observed that the correlation tends to increase with the increase in leverage values, for example, in Tables-6.a.1 to 6.a.6, the values of r_{12}, are 0.59185, 0.79995, 0.88620, 0.92744, 0.97241 and 0.98300 respectively when the value of n = 20. Similar results are observed for different sample sizes. Throughout the simulations we observe that the performance of Huber-1 method is very poor. The performance of each of the other methods is good.

6.8.3 Simulation Results for 10% equal High Leverage Point

In this subsection 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, X_2 and X_3 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 Tables-6.b.1 to 6.b.6 those are based on the average of 10,000 simulations.

Table 6.b.1: Simulation results in presence of 10% equal high leverage point = 2

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.74840	0.73694	0.73566	0.73238	0.73204	0.73077
	2 Mean	0.03582	0.01212	0.01884	0.00637	0.01550	0.01037
	3 Mean	0.74870	0.73714	0.73572	0.73240	0.73207	0.73077
	Huber-1	0.029200	0.08210	0.01884	0.73307	0.73204	0.73077
	Huber-2	0.75010	0.73694	0.73566	0.73238	0.73204	0.73077
	Potential (mean)	0.74212	0.68194	0.67850	0.65638	0.66304	0.66624
	Potential (med)	0.08620	0.017290	0.28270	0.34580	0.56557	0.68528
	G.P.	0.01597	-0.00184	0.00329	-0.00850	0.00236	0.00010
r_{23}	ACTUAL	0.74435	0.73927	0.73699	0.73415	0.73108	0.73084
	2 Mean	0.02084	0.03281	0.02185	0.00904	0.01083	0.01059
	3 Mean	0.74453	0.73946	0.73717	0.73422	0.73113	0.73084
	Huber-1	0.28476	0.09508	0.02185	0.73476	0.73108	0.73084
	Huber-2	0.74589	0.73927	0.73699	0.73415	0.73108	0.73084
	Potential (mean)	0.73706	0.68546	0.67916	0.65662	0.66024	0.66641
	Potential (med)	0.0718	0.18220	0.28370	0.34520	0.56510	0.68644
	G.P.	0.00214	0.00856	0.00248	-0.00419	-0.00233	0.00063
r_{23}	ACTUAL	0.74721	0.73770	0.74034	0.73474	0.73351	0.72969
	2 Mean	0.02942	0.02061	0.03109	0.01186	0.02120	0.00518
	3 Mean	0.74761	0.73798	0.74046	0.73479	0.73353	0.72969
	Huber-1	0.29060	0.08304	0.03109	0.73535	0.73351	0.72969
	Huber-2	0.74936	0.73770	0.74034	0.73474	0.73351	0.72969
	Potential (mean)	0.74001	0.68372	0.68227	0.66001	0.66363	0.66596
	Potential (med)	0.07770	0.17640	0.29280	0.34890	0.56807	0.68458
	G.P.	0.00858	0.00419	0.01574	-0.00175	0.00666	-0.00403

Table 6.b.2: Simulation results in presence of 10% equal high leverage point = 3

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.88876	0.88645	0.88618	0.88442	0.88375	0.88296
	2 Mean	0.01626	0.01596	0.02401	0.01124	0.00880	0.01012
	3 Mean	0.88883	0.88652	0.88627	0.88445	0.88376	0.88296
	Huber-1	0.27753	0.07861	0.02401	0.88488	0.88375	0.88296
	Huber-2	0.88955	0.88645	0.88618	0.88442	0.88375	0.88296
	Potential (mean)	0.88808	0.84647	0.83593	0.83091	0.83162	0.84025
	Potential (med)	0.04012	0.08693	0.15510	0.20530	0.44350	0.66750
	G.P.	-0.00126	-0.00077	0.00797	-0.00382	-0.00225	0.00117
r_{23}	ACTUAL	0.88803	0.88511	0.88583	0.88463	0.88393	0.88272
	2 Mean	0.01209	0.00756	0.02012	0.01359	0.01222	0.00960
	3 Mean	0.88810	0.88516	0.88586	0.88467	0.88395	0.88272
	Huber-1	0.27518	0.07518	0.02012	0.88503	0.88393	0.88272
	Huber-2	0.88872	0.88511	0.88583	0.88463	0.88393	0.88272
	Potential (mean)	0.88737	0.84522	0.83550	0.83000	0.83086	0.84133
	Potential (med)	0.03127	0.07710	0.15040	0.21170	0.44280	0.66810
	G.P.	-0.00513	-0.00674	0.00527	-0.00052	0.00014	-0.00007
r_{23}	ACTUAL	0.88896	0.88684	0.88459	0.88458	0.88308	0.88329
	2 Mean	0.02029	0.01790	0.01050	0.01524	0.00621	0.01269
	3 Mean	0.88897	0.88692	0.88471	0.88464	0.88308	0.88329
	Huber-1	0.27885	0.07880	0.01050	0.88499	0.88308	0.88329
	Huber-2	0.88957	0.88684	0.88459	0.88458	0.88308	0.88329
	Potential (mean)	0.88772	0.84804	0.83185	0.83439	0.82940	0.84248
	Potential (med)	0.03770	0.08920	0.13910	0.20850	0.43730	0.66800
	G.P.	-0.00018	0.00179	-0.00530	-0.00075	-0.00603	0.00300

Table 6.b.3: Simulation results in presence of 10% equal high leverage point = 4

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.94105	0.93889	0.93797	0.93777	0.93737	0.93656
	2 Mean	0.03514	0.02915	0.01330	0.01528	0.01661	0.00691
	3 Mean	0.94109	0.93893	0.93802	0.93779	0.93738	0.93656
	Huber-1	0.29283	0.09354	0.01330	0.93805	0.93737	0.93656
	Huber-2	0.94145	0.93889	0.93797	0.93777	0.93737	0.93656
	Potential (mean)	0.93997	0.91205	0.88869	0.89423	0.89697	0.91311
	Potential (med)	0.05029	0.07401	0.09700	0.18940	0.40650	0.61940
	G.P.	0.01800	0.00722	-0.00470	0.00051	0.00561	-0.00187
r_{23}	ACTUAL	0.93980	0.93813	0.93783	0.93760	0.93719	0.93662
	2 Mean	0.02160	0.01265	0.01194	0.01447	0.01349	0.01080
	3 Mean	0.93984	0.93816	0.93788	0.93763	0.93720	0.93662
	Huber-1	0.28377	0.07964	0.01194	0.93784	0.93719	0.93662
	Huber-2	0.94023	0.93813	0.93783	0.93760	0.93719	0.93662
	Potential (mean)	0.93908	0.91180	0.88835	0.89510	0.89773	0.91308
	Potential (med)	0.03471	0.06202	0.09630	0.18270	0.40530	0.61970
	G.P.	0.00110	-0.00044	-0.00398	-0.00067	0.00115	0.00083
r_{23}	ACTUAL	0.93959	0.93833	0.93792	0.93753	0.93686	0.93654
	2 Mean	0.01772	0.01538	0.01313	0.01450	0.00785	0.00891
	3 Mean	0.93965	0.93838	0.93793	0.93755	0.93687	0.93654
	Huber-1	0.28051	0.08290	0.01313	0.936778	0.93686	0.93654
	Huber-2	0.93996	0.93833	0.93792	0.93753	0.93686	0.93654
	Potential (mean)	0.93879	0.91181	0.88727	0.89448	0.89722	0.91290
	Potential (med)	0.03233	0.06403	0.09760	0.18760	0.40270	0.61850
	G.P.	-0.00132	-0.00215	-0.00315	0.00030	-0.00399	-0.00153

Table 6.b.4: Simulation results in presence of 10% equal high leverage point = 5

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.96295	0.96214	0.96155	0.96135	0.96095	0.96063
	2 Mean	0.02998	0.02013	0.01150	0.01694	0.01219	0.00703
	3 Mean	0.96296	0.96216	0.96160	0.96136	0.96095	0.96063
	Huber-1	0.29452	0.09122	0.01150	0.96151	0.96095	0.96063
	Huber-2	0.96320	0.96215	0.96155	0.96135	0.96095	0.96063
	Potential (mean)	0.96295	0.94133	0.91687	0.91959	0.92453	0.93436
	Potential (med)	0.04562	0.06413	0.11330	0.18510	0.38490	0.61830
	G.P.	0.00898	0.00253	-0.00465	0.00122	-0.00129	-0.00198
r_{23}	ACTUAL	0.96311	0.96136	0.96170	0.96097	0.96072	0.96060
	2 Mean	0.02534	0.00695	0.01382	0.00458	0.00649	0.00890
	3 Mean	0.96315	0.96140	0.96172	0.96097	0.96072	0.96060
	Huber-1	0.28287	0.06773	0.01382	0.96110	0.96072	0.96060
	Huber-2	0.96339	0.96137	0.96170	0.96097	0.96072	0.96060
	Potential (mean)	0.96313	0.93990	0.91781	0.91932	0.92396	0.93388
	Potential (med)	0.04497	0.05208	0.11640	0.17520	0.38110	0.61870
	G.P.	0.00761	-0.00963	-0.00167	-0.01070	-0.00631	-0.00126
r_{23}	ACTUAL	0.96264	0.96185	0.96165	0.96123	0.96102	0.96055
	2 Mean	0.01824	0.01707	0.01382	0.00976	0.01224	0.00650
	3 Mean	0.96267	0.96190	0.96169	0.96123	0.96102	0.96055
	Huber-1	0.28235	0.08147	0.01382	0.96137	0.96102	0.96055
	Huber-2	0.96291	0.96186	0.96165	0.96123	0.96102	0.96055
	Potential (mean)	0.96265	0.94131	0.91836	0.92109	0.92404	0.93444
	Potential (med)	0.03039	0.06092	0.11290	0.18000	0.38750	0.61910
	G.P.	-0.00185	-0.00005	-0.00264	-0.00260	0.00161	-0.00163

Table 6.b.5: Simulation results in presence of 10% equal high leverage point = 8

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.98611	0.98593	0.98578	0.98585	0.98559	0.98543
	2 Mean	0.01916	0.01573	0.02121	0.02329	0.01504	0.00691
	3 Mean	0.98611	0.98495	0.98580	0.98585	0.98560	0.98543
	Huber-1	0.27609	0.08205	0.02121	0.98590	0.98559	0.98543
	Huber-2	0.98617	0.98593	0.98578	0.98585	0.98559	0.98543
	Potential (mean)	0.98611	0.96250	0.95458	0.94815	0.94800	0.96682
	Potential (med)	0.02881	0.05028	0.10640	0.17450	0.37830	0.57660
	G.P.	-0.00057	0.00139	0.00305	0.01043	0.00379	-0.00223
r_{23}	ACTUAL	0.98603	0.98591	0.98579	0.98569	0.98565	0.98556
	2 Mean	0.01561	0.01675	0.02014	0.01497	0.01932	0.01521
	3 Mean	0.98604	0.98592	0.98581	0.98569	0.98565	0.98556
	Huber-1	0.26550	0.08588	0.02014	0.98575	0.98565	0.98556
	Huber-2	0.98610	0.98591	0.98579	0.98569	0.98565	0.98556
	Potential (mean)	0.98604	0.96191	0.95429	0.94933	0.94755	0.96668
	Potential (med)	0.02690	0.05006	0.10280	0.17110	0.38010	0.58100
	G.P.	-0.00150	-0.00060	0.00258	0.00284	0.00588	0.00408
r_{23}	ACTUAL	0.98626	0.985963	0.98573	0.98559	0.98565	0.98546
	2 Mean	0.02298	0.01843	0.01725	0.00822	0.01713	0.00863
	3 Mean	0.98627	0.98593	0.98575	0.98559	0.98565	0.98546
	Huber-1	0.29015	0.07426	0.01725	0.98563	0.98565	0.98546
	Huber-2	0.98631	0.98593	0.98573	0.98559	0.98565	0.98546
	Potential (mean)	0.98627	0.96259	0.95361	0.94792	0.94622	0.96672
	Potential (med)	0.03446	0.05260	0.10530	0.16640	0.37810	0.57910
	G.P.	0.00443	-0.00259	-0.00022	-0.00803	0.00516	-0.00031

Table 6.b.6: Simulation results in presence of 10% equal high leverage point = 10

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.99147	0.99123	0.99103	0.99110	0.99093	0.99089
	2 Mean	0.03326	0.01806	0.00914	0.02081	0.01054	0.00904
	3 Mean	0.99148	0.99123	0.99103	0.99110	0.99093	0.99089
	Huber-1	0.29059	0.08454	0.00914	0.99114	0.99093	0.99089
	Huber-2	0.99155	0.99123	0.99103	0.99110	0.99093	0.99089
	Potential (mean)	0.99147	0.96924	0.95036	0.94607	0.95539	0.96990
	Potential (med)	0.01121	0.04971	0.09560	0.16120	0.36600	0.57880
	G.P.	0.01172	0.00298	-0.00708	0.00464	-0.00141	-0.00027
r_{23}	ACTUAL	0.99149	0.99129	0.99116	0.99107	0.99096	0.99093
	2 Mean	0.02277	0.02644	0.02039	0.01687	0.01393	0.01488
	3 Mean	0.99150	0.99130	0.99116	0.99108	0.99097	0.99093
	Huber-1	0.28703	0.09350	0.02039	0.99112	0.99096	0.99093
	Huber-2	0.99155	0.99129	0.99116	0.99107	0.99096	0.99093
	Potential (mean)	0.99147	0.96662	0.95107	0.94588	0.95720	0.96970
	Potential (med)	0.03719	0.05385	0.10840	0.15360	0.36910	0.57730
	G.P.	0.00315	0.00902	0.00496	0.00218	0.00299	0.00331
r_{23}	ACTUAL	0.99139	0.99110	0.99106	0.99104	0.99097	0.99088
	2 Mean	0.02117	0.01023	0.01549	0.01140	0.01415	0.00718
	3 Mean	0.99140	0.99110	0.99107	0.99104	0.99097	0.99088
	Huber-1	0.28090	0.07426	0.01549	0.99108	0.99097	0.99088
	Huber-2	0.99145	0.99110	0.99106	0.99104	0.99097	0.99088
	Potential (mean)	0.99139	0.96680	0.94854	0.94642	0.95554	0.97034
	Potential (med)	0.03823	0.03849	0.10500	0.15100	0.36890	0.57650
	G.P.	0.00088	-0.00913	-0.00075	-0.00140	0.00098	-0.00165

6.8.4 Simulation Results Discussion for 10% Equal High Leverage Point

We observe in Tables 6.b.1 to 6.b.6 that for every n, the presence of multiple high leverage points causes strong multicollinearity. Here it is also observed that the correlations between the X's tend to reduce slightly with the increase in sample size for example, in Table-6.b.1, the values of r_{12}, r_{13}, and r_{23} are (0.74840, 0.74435, 0.74721), (0.73694, 0.73927, 0.73770), (0.73566, 0.73699, 0.74034), (0.73238, 0.73415, 0.73474), (0.73204, 0.73108, 0.73351) and (0.73077, 0.73084, 0.72969) for n=20, 30, 40, 50, 100 and 200 respectively. In Tables-6.b.1 to 6.b.6, the values of r_{12} are 0.74840, 0.88876, 0.94105, 0.96295, 0.98611and 0.99147 respectively for the values of x=2, 4, 6, 8, 10, 12,..... with n=20. Similar results are obtained for different sample sizes and for the other two correlation values r_{13} and r_{23}. This implies that the correlation between the X's 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-1 method is appeared to be good for small samples, but even for the moderate sample size like 50 it breaks down. Potential (med) 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 generalized potentials performed best.

6.8.5 Simulation Results Discussion for 10% Unequal High Leverage Point

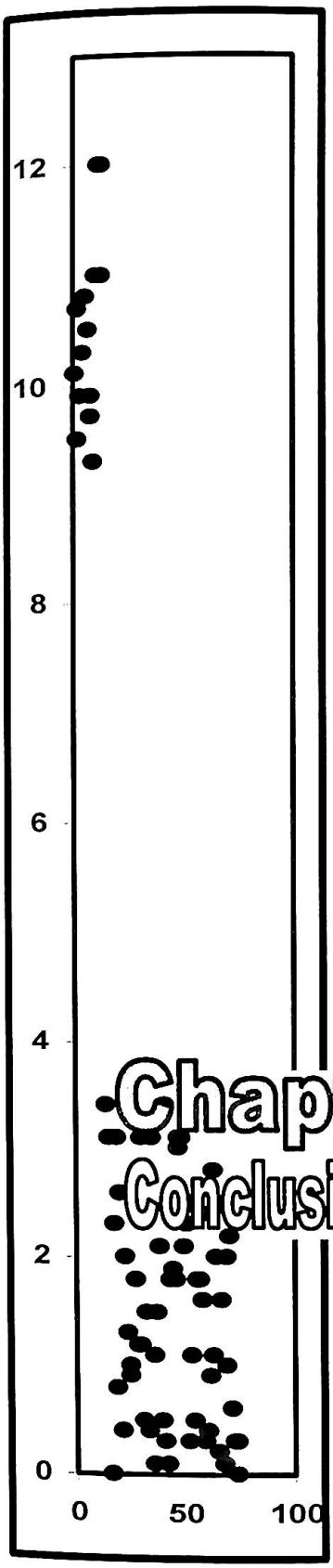
In this subsection first we present the simulation results where the X variables contain 10% high leverage points having unequal weights. For each of the cases the first 90% observations are simulated as Uniform (0, 1). The last 10% observations of X_1 and X_2 ands X_3 are taken serially from a set of observations starting from 2 and then having increments of 2 (i.e. x=2, 4, 6, 8, 10,12, ...) 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 presented in Tables-7.c those are based on the average of 10,000 simulations.

Table 6.c.1: Simulation results in presence of 10% unequal high leverage point (2, 3, 4, 5,,...., 40)

Correlation	Measures	Value of n					
		20	30	40	50	100	200
r_{12}	ACTUAL	0.9416	0.94983	0.96983	0.97992	0.99445	0.99855
	2 Mean	0.62036	0.51774	0.83446	0.92516	0.96257	0.98960
	3 Mean	0.60657	0.86308	0.93558	0.92173	0.97569	0.99496
	Huber-1	0.81202	0.58111	0.83446	0.92210	0.98956	0.99855
	Huber-2	0.60744	0.86302	0.93555	0.97992	0.99445	0.99855
	Potential (mean)	0.60612	0.86302	0.93555	0.96318	0.98480	0.99650
	Potential (med)	0.63702	0.67554	0.83419	0.90944	0.96743	0.99203
	G.P.	0.00557	0.00408	0.00676	-0.00291	0.00160	0.00146
r_{23}	ACTUAL	0.90398	0.94964	0.96984	0.98010	0.99441	0.99854
	2 Mean	0.61798	0.51465	0.83386	0.92618	0.96241	0.98955
	3 Mean	0.60490	0.86243	0.93560	0.92243	0.97644	0.99494
	Huber-1	0.80794	0.57908	0.83386	0.92287	0.98948	0.99854
	Huber-2	0.60667	0.86243	0.93559	0.98010	0.99441	0.99854
	Potential (mean)	0.60474	0.86243	0.93559	0.96351	0.98470	0.99649
	Potential (med)	0.63501	0.67343	0.83351	0.91037	0.96722	0.99200
	G.P.	-0.00228	-0.00208	0.00912	0.00112	-0.00177	-0.00461
r_{23}	ACTUAL	0.90229	0.94966	0.96984	0.98021	0.99443	0.99854
	2 Mean	0.61284	0.51571	0.83441	0.92642	0.96235	0.98958
	3 Mean	0.59845	0.86254	0.93559	0.92291	0.97648	0.99495
	Huber-1	0.80540	0.57898	0.83441	0.92328	0.98950	0.99854
	Huber-2	0.60043	0.86250	0.93558	0.98021	0.99443	0.99854
	Potential (mean)	0.59819	0.86250	0.93558	0.96372	0.98473	0.99650
	Potential (med)	0.62842	0.67782	0.83426	0.91115	0.96723	0.99203
	G.P.	-0.01444	-0.00254	0.00540	0.00474	-0.00318	0.00113

6.8.6 Simulation Results Discussion for 10% Unequal High Leverage Point

We observe from results of Table-6.c 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 increase 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 generalized 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 generalized potentials is quite outstanding. For all samples we observe that the omission of the cases identified by this method produces very low correlation coefficients can entirely remove the multicollinearity effect from the data.



Chapter Seven

Conclusion and Areas of Further Research

Chapter Seven

Conclusion and Areas of Further Research

7.1 Discussion of Results

In our study, we considered some commonly used leverage measures in a comparative study to investigate their sensitivity and usefulness in the detection of high leverage cases under a variety of situations. We notice two major things in our study. Firstly, many of the commonly used diagnostics are too much sensitive, i.e, they have a tendency to identify good cases as high leverage points and secondly, all of them more or less suffer from masking in the presence of multiple high leverage points. Hoaglin and Welsch's twice-the-mean rule performed best overall. This method is very effective in the identification of a single high leverage point and in a situation where multiple high leverage cases have equal weights. When multiple high leverage cases have unequal weights, its performance is not entirely satisfactory but it is still better than any other methods considered in this

paper. This method also produceed a low swamping effect which indicates that twice-the-mean rule is not too prone to declare low leverage cases as points of high leverage. This method is followed by Hadi's Potential (median) whose overall performance can be considered as satisfactory. The conservative suggestion of Huber (Huber-1 method) is the next choice. This method is successful in identifying a single high leverage point, but when a group of high leverage points is present it becomes successful only when the sample size is small. But for small sample we observe that this method possesses a very high swamping rate in every occasion. The good thing about the other three methods considered in this paper that neither of them is sensitive in no high leverage situations and they are successful in the identification of a single high leverage case but their performances seemed to be very poor in the presence of multiple high leverage cases.

The ineffectiveness of all of the commonly used diagnostic measures in the identification of multiple high leverage cases tells us that we require new diagnostics which are designed to identify multiple high leverage points.

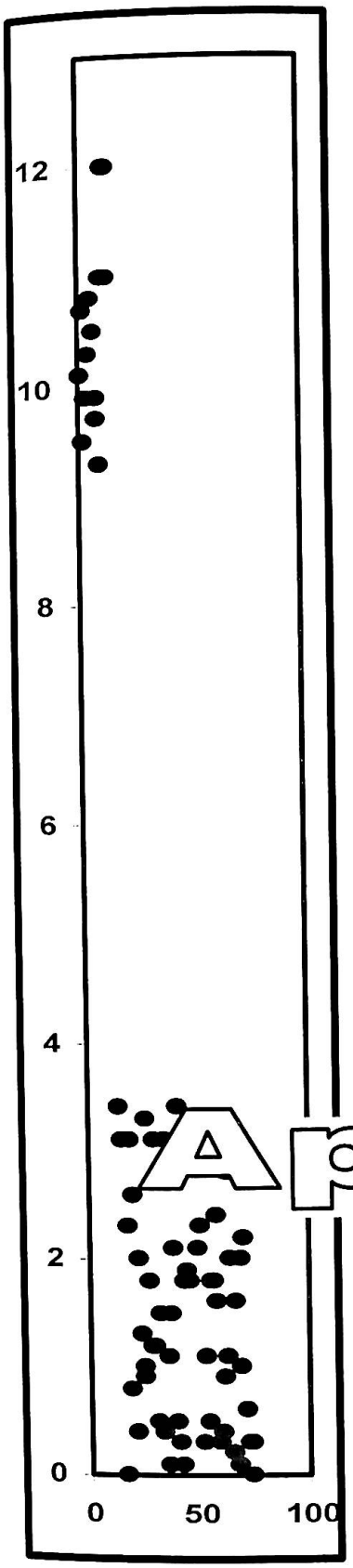
Then we introduced a group deleted version of Hadi's potential and follow Imon (1996) to call it as generalized potentials. We proposed a new technique based on generalized potentials to identify multiple high leverage points. At first we consider a well known data set and observe that this newly proposed diagnostic becomes very successful in the identification of multiple high leverage points when all other commonly used diagnostics fail to do so. The simulation results also support the merit of using our proposed diagnostics. It possesses a relatively low swamping rate in a now high leverage situation, but it produced outstanding result when the data contains several high leverage points.

In our study, we also introduce a new graphical display for location multiple high leverage points with outliers and influential observations. We observe from a well known data set that our proposed diagnostic plot could be very effective in the detection of multiple unusual observations.

We also observe in our study that high leverage point may produce strong multicollinearity among the regressors. The omission of the high leverage points could be an option to get rid of this problem. Since the traditional detection procedures are in doubt, we observe that neither of them becomes successful to remove the problem of multicollinearity. But the diagnostic technique based on generalized potentials as we suggest in our study produces stunning results. The omission of observations identified by this method can remove the effect of multicollinearity entirely from the data.

7.2 Areas of Further Research

In our study, we only considered a linear regression setup. But in many occasion we have to consider non-linear regression model and we sincerely believe that our ideas could be extended to non-linear regression. We anticipate that many of our results could be applied readily to logistic regression analysis which is a growing area of research. We also believe that the group deleted leverage measures could be successfully used to measure influence of observation.



Appendix

Appendix

MINITAB Programs

Program-1: MINITAB Simulation Program for Identification of Multiple High Leverage Points When 10% Unequal High Leverage Point are Present in the data set for a Sample of Size 100:

```
Let k40 = k40+1
rand 45 c2-c4;
uniform.
let c2 (91) = 2
let c2 (92) = 4
let c2 (93) = 6
let c2 (94) = 8
```

```
let c2 (95) = 10
let c2 (96) = 12
let c2 (97) = 14
let c2 (98) = 16
let c2 (99) = 18
let c2 (100) = 20
let c3 (91) = 2
let c3 (92) = 4
let c3 (93) = 6
let c3 (94) = 8
let c3 (95) = 10
let c3 (96) = 12
let c3 (97) = 14
let c3 (98) = 16
let c3 (99) = 18
let c3 (100) = 20
let c4 (91) = 2
let c4 (92) = 4
let c4 (93) = 6
let c4 (94) = 8
let c4 (95) = 10
let c4 (96) = 12
let c4 (97) = 14
let c4 (98) = 16
let c4 (99) = 18
let c4 (100) = 20
rand 100 c5
let c1 = 1-0.5*c2+3*c3-2*c4+c5
regr c1 3 c2-c4;
hi c6.
```

```
let c7 = c6/(1-c6)
let c8 = c6 > (8/100)
let c9 = c6 > (12/100)
let c10 = c6 > 0.2
let c11 = c6 > 0.5
let k1 = mean (c7)
let k2 = stdev (c7)
let k3 = k1+ (3*k2)
let k4 = median(c7)
let c12 = c7-k4
let c13 = abso (c12)
let k5 = median (c13)
let k6 = k4 + (5*k5)
let k7 = k4+(10*k5)
let c14 = c7 > k3
let c15 = c7> k6
let c16 = c7>k7
set c17
1 (1: 1 / 1) 100
copy c17 c2 c3 c4 m1
tran m1 m2
copy c2-c4 c21-c23;
omit 81 : 100.
Set c18
(1 : 1 / 1) 80
copy c18 c21 c22 c23 m3
tran m3 m4
mult m4 m3 m5
inverse m5 m6
mult m1 m6 m7
```

```
mult m7 m2 m8
diag m8 c24
let c25 = c24-median (c24)
let c26 = abso (c25)
let k8 = median (c26) / 0.6745
let k9 = median (c24)+3*k8
let k10 = median (c24)+5*k8
let c27 = c24 > k9
let c28 = c24 > k10
set c29
80 (0)
20 (1)
end
let c30 = (1- c29)
let c31 = c8*c29
let c32 = c8*c30
let c33 = c9*c29
let c34 = c9*c30
let c35 = c10*c29
let c36 = c10*c30
let c37 = c11*c29
let c38 = c11*c30
let c39 = c14*c29
let c40 = c14*c30
let c41 = cc15*c29
let c42 = c15*c30
let c43 = c16*c29
let c44 = c16*c30
let c45 = c27*c29
let c46 = c27*c29
```

```
let c46 = c28*c29
let c47 = c28*c30
let k11 = sum (c31)
let k12 = sum (c32)
let k13 = sum (c33)
let k14 = sum (c34)
let k15 = sum (c35)
let k16 = sum (c36)
let k17 = sum (c37)
let k18 = sum (c38)
let k19 = sum (c39)
let k20 = sum (c40)
let k21 = sum (c41)
let k22 = sum (c42)
let k23 = sum (c43)
let k24 = sum (c44)
let k25 = sum (c45)
let k26 = sum (c46)
let k27 = sum (c47)
let k28 = sum (c48)
let c50 (k40) = k11
let c51 (k40) = k12
let c52 (k40) = k13
let c53 (k40) = k14
let c54 (k40) = k15
let c55 (k40) = k16
let c56 (k40) = k17
let c57 (k40) = k18
let c58 (k40) = k19
let c59 (k40) = k20
```

```
let c60 (k40) = k21
let c61 (k40) = k22
let c62 (k40) = k23
let c63 (k40) = k24
let c64 (k40) = k25
let c65 (k40) = k26
let c66 (k40) = k27
let c67 (k40) = k28
name c50 '2Mi'
name c51 '2Ms'
name c52 '3Mi'
name c53 '3Ms'
name c54 'Hu1i'
name c55 'Hu1s'
name c56 'Hu2i'
name c57 'Hu2s'
name c58 'H.mean i'
name c59 'H.mean s'
name c60 'H.med (c=5) i'
name c61 'H.med (c=5) s'
name c62 'H.med (c=10) i'
name c63 'H.med (c=10) s'
name c64 'GP (c=3) i'
name c65 'GP (c=3) s'
name c66 'GP (c=5) i'
name c67 'GP (c=3) s'
end
```

Program-2: MINITAB Simulation Program to Investigate How High Leverage Points Behave as a Source of Multicollinearity When 10% Unequal High Leverage Point are Present in the data set for a Sample of Size 100:

```
Let k40 = k40+1
rand 45 c2-c4;
uniform.
let c2 (91) = 2
let c2 (92) = 4
let c2 (93) = 6
let c2 (94) = 8
let c2 (95) = 10
let c2 (96) = 12
let c2 (97) = 14
let c2 (98) = 16
let c2 (99) = 18
let c2 (100) = 20
let c3 (91) = 2
let c3 (92) = 4
let c3 (93) = 6
let c3 (94) = 8
let c3 (95) = 10
let c3 (96) = 12
let c3 (97) = 14
let c3 (98) = 16
let c3 (99) = 18
let c3 (100) = 20
let c4 (91) = 2
```

```
let c4 (92) = 4
let c4 (93) = 6
let c4 (94) = 8
let c4 (95) = 10
let c4 (96) = 12
let c4 (97) = 14
let c4 (98) = 16
let c4 (99) = 18
let c4 (100) = 20
rand 100 c5
let c1 = 1-0.5*c2+3*c3-2*c4+c5
regr c1 3 c2-c4;
```

h_i c6.

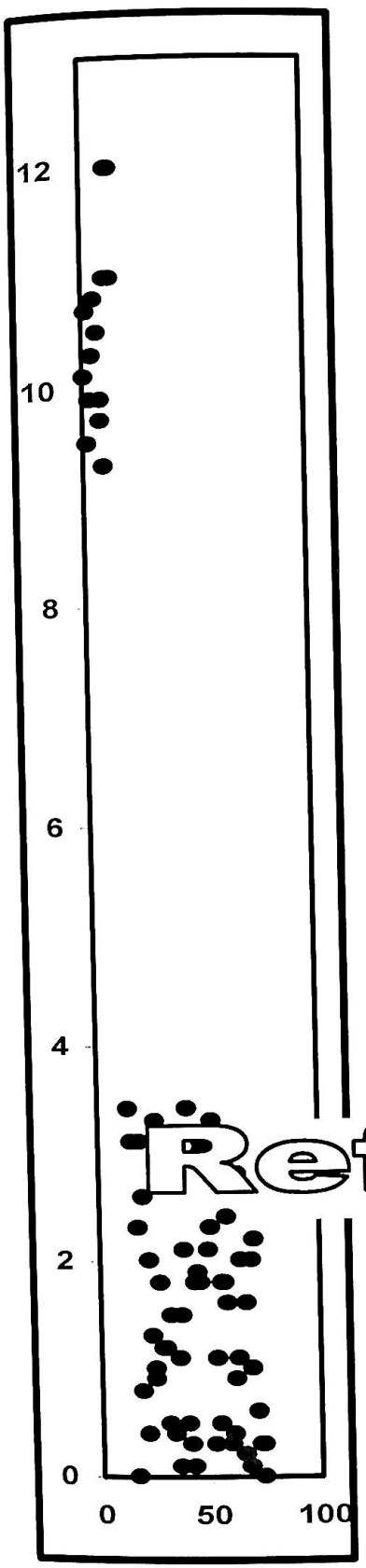
```
let c7 = c6/(1-c6)
let c8 = c6 > (8/100)
let c9 = c6 > (12/100)
let c10 = c6 > 0.2
let c11 = c6 > 0.5
let k1 = mean (c7)
let k2 = stdev (c7)
let k3 = k1+ (3*k2)
let k4 = median(c7)
let c12 = c7-k4
let c13 = abso (c12)
let k5 = median (c13)
let k6 = k4 + (5*k5)
let c14 = c7 > k3
let c15 = c7> k6
set c17
1 (1: 1 / 1) 100
copy c17 c2 c3 c4 m1
```

```
tran m1 m2
copy c2-c4 c21-c23;
omit 81 : 100.
Set c18
(1 : 1 / 1) 80
copy c18 c21 c22 c23 m3
tran m3 m4
mult m4 m3 m5
inverse m5 m6
mult m1 m6 m7
mult m7 m2 m8
diag m8 c24
let c25 = c24-median (c24)
let c26 = abso (c25)
let k8 = median (c26) / 0.6745
let k9 = median (c24)+3*k8
let c27 = c24 > k9
copy c2-c4 c30-c32;
omit c8 = 1.
copy c2-c4 c33-c35;
omit c9 = 1.
copy c2-c4 c36-c38;
omit c10 = 1.
copy c2-c4 c39-c41;
omit c11 = 1.
copy c2-c4 c42-c44;
omit c14 = 1.
copy c2-c4 c45-c47;
omit c15 = 1.
copy c2-c4 c48-c50;
omit c27 = 1.
```

```
corr c2-c4 m10
copy m10 c51-c53
let k10 = c51 (2)
let k11 = c51 (3)
let k12 = c52 (3)
corr c30-c32 m11
copy m11 c54-c56
let k13 = c54 (2)
let k14 = c54 (3)
let k15 = c55 (3)
corr c33-c35 m12
copy m12 c57-c59
let k16 = c57 (2)
let k17 = c57 (3)
let k18 = c58 (3)
corr c36-c38 m13
copy m13 c60-c62
let k19 = c60 (2)
let k20 = c60 (3)
let k21 = c61 (3)
corr c39-c41 m14
copy m14 c63-c65
let k22 = c63 (2)
let k23 = c63 (3)
let k24 = c64 (3)
corr c42-c44 m15
copy m15 c66-c68
let k25 = c66 (2)
let k26 = c66 (3)
let k27 = c67 (3)
corr c45-c47 m16
```

```
copy m16 c69-c71
let k28 = c69 (2)
let k29 = c69 (3)
let k30 = c70 (3)
corr c48-c50 m17
copy m17 c72-c74
let k31 = c72 (2)
let k32 = c72 (3)
let k33 = c73 (3)
let c75 (k40) = k10
let c76 (k40) = k11
let c77 (k40) = k12
let c78 (k40) = k13
let c79 (k40) = k14
let c80 (k40) = k15
let c81 (k40) = k16
let c82 (k40) = k17
let c83 (k40) = k18
let c84 (k40) = k19
let c85 (k40) = k20
let c86 (k40) = k21
let c87 (k40) = k22
let c88 (k40) = k23
let c89 (k40) = k24
let c90 (k40) = k25
let c91 (k40) = k26
let c92 (k40) = k27
let c93 (k40) = k28
let c94 (k40) = k29
let c95 (k40) = k30
let c96 (k40) = k31
```

```
let  c97 (k40) = k32
let  c98 (k40) = k33
name  c75  'corr (1, 2)
name  c76  'corr (1, 3)
name  c77  'corr (2, 3)
name  c78  '2Mcorr (1, 2)
name  c79  '2Mcorr (1, 3)
name  c80  '2Mcorr (2, 3)
name  c81  '3Mcorr (1, 2)
name  c82  '3Mcorr (1, 3)
name  c83  '3Mcorr (2, 3)
name  c84  'Hu1 corr (1, 2)
name  c85  'Hu1 corr (1, 3)
name  c86  'Hu1 corr (2, 3)
name  c87  'Hu2 corr (1, 2)
name  c88  'Hu2 corr (1, 3)
name  c89  'Hu2 corr (2, 3)
name  c90  'H mean corr (1, 2)
name  c91  'H mean corr (1, 3)
name  c92  'H mean corr (2, 3)
name  c93  'H med corr (1, 2)
name  c94  'H med corr (1, 3)
name  c95  'H med corr (2, 3)
name  c96  'GP corr (1, 2)
name  c97  'GP corr (1, 3)
name  c98  'GP corr (2, 3)
end
```



References

References

[1] Andrews, D.F. and Pregibon, D. (1978), "Finding of Outliers that Matter," *Journal of the Royal Statistical Society*, Series B, **40**, 85-93.

[2] Andrews, D.F., Bickel, P.J., Hampel, F.W., Huber, P.J., Rogers, W.H.,and Tukey, J.W. (1972), *Robust Estimations of Location: Survey and Advances.* Princeton University Press. New Jersey.

[3] Arnold, S.F. (1980), "Asymptotic Validity of F Test for the Ordinary Linear Model and the Multiple Correlation Model," *Journal of the American Statistical Association*, **75**, 890-894.

[4] Atkinson, A.C. (1985), *Plots, Transformations and Regression*, Oxford University Press, Oxford.

[5] Barnett, V., and Lewis, T. (1994), *Outlier in Statistical Data*, 3rd. Ed., Wiley, New York.

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

[7] Bera, A.K. and Jarque, C.M. (1982), "Model Specification Tests: A Simultaneous Approach," *Journal of Econometrics*, **20**, 59-82.

[8] Chatterjee, S. and Hadi, A.S. (1986), "Influential Observations, High Leverage Points, and Outliers in Linear Regression (With Diccussion)."*Statistical Science*, **1**, 379-416.

[9] Chatterjee, S. and Hadi, A.S. (1988), *Sensitivity Analysis in Linear Regression*, Wiley, New York.

[10] Cook, R.D. and Hawkins, D.M. (1990). "Comment on 'Unmasking Multivariate Outliers and Leverage Points' by Rousseeuw, P.J. and van Zomeren, B.C *Journal of the American Statistical Association,* **85**, 640-644.

[11] Farrar, D.E. and Glauber, R.R.(1967), "Multicollinearity in Regression Analysis: The problem revisited," *Review of Economics Statistics*, **49**, 92-107.

[12] Frisch, F. (1934), *Statistical Confluence analysis by Mean of Complete Regression System*, University Institute of Economics, Oslo, publication 5.

[13] Galton F. (1886), "Family Likeness in Stature," *Journal of the Royal Statistical Society,* **40**, 42-72.

[14] Geary, R.C. (1947), "Testing for Normality," *Biometrika*, **34**, 209-242.

[15] Gentle, J.E. (1982), '*Monte Carlo Methods*' *in Encyclopedia of Statistical Scince*, Kotz, S. and Johnson N.L. (Ed) Wiley, New York.

[16] Gentleman, J.F. and Wilk, M.B. (1975), "Detecting Outliers in a Two-way Table:II Supplementing the Direct Analysis of Residuals" *Biometrics*, **31**, 387-410.

[17] Ghosh, S. (1996). "A New Graphical Tool to Detect non-Normality". *Journal of the Royal Statistical Society*, Series B, **58**, 691-702.

[18] Gnanadesikan, R. (1977), *Methods for Statistical Analysis of Multivariate Data,* Wiley, New York.

[19] Gray, J.B. (1986). A Simple Graphic for Assessing Influence in Regression. *Journal of Statistical Computation and Simulation.* **24**, 121-134.

[20] Gujarati, D. N. (1995), *Econometric Methods,* Mc Grow-Hill, Inc.

[21] Gunst, R. F. and Mason, R. L. (1977), "Biased Estimation in Regression: An Evaluation Using Mean Squared Error," *Journal of the American Statistical Association*, **72**, 616-628.

[22] Hadi, A.S.(1992), "A New Measure of Overall Potential Influence in Linear Regression," *Computational Statistics and Data Analysis*, **14**, 1-27.

[23] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, P.J., (1986), *Robust Statistics: The Approach Based on Influence Function*, Wiley, New York.

[24] Hartley, H.O. (1977), *Statistical Methods for Digital Computers*, Wiley, New York.

[25] 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.

[26] Henderson, H.V. and Searle, S.R. (1981), On Deriving the Inverse of a Sum of Matrices, *SIAM Review*, **22**, 53-60.

[27] Hoaglin, D.C., and Welsch, R.E. (1978), "The Hat Matrix in Regression and ANOVA," *Journal of the American Statistical Association*, **32**, 17-22.

[28] Hocking, R.R., and Pendletion, O.J. (1983), "The Regression Dilemma," *Communications in Statistics*. Series A, **12**, 497-527.

[29] Hoerl, A. E. and Kennard R.W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, **12**, 55-67.

[30] Huber, P.J. (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *The Annals of Statistics*, **1**, 799-821.

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

[32] 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.*

[33] Imon, A.H.M.R. (1999), "On PRESS Residuals," *Journal of Statistical Research*, **33**, 59-65.

[34] Imon, A.H.M.R. (2000), "Recent Computational Advances in Linear Regression," Proceedings of the 7-*th* Annual Conference of Bangladesh Statistical Association, pp. 127-134.

[35] Imon, A.H.M.R. (2003), "Regression Residual, Moments and Their Use in Tests for Normality," *Communucations in Statistics-Theory and Methods,* **32**, 1021-1034.

[36] Judge, G.G., Griffith, W.E.,Hill, R.C., Lutkepohl, H., and Lee, T. (1985), *Theory and Practice of Econometrics*, 2nd Ed., Wiley New York.

[37] Kadane, J.B. (Ed) (1984), *Robustness in Bayesian Analysis*, Elsevier North-Holland, Amsterdam.

[38] Kendal, M.G. and Buckland, W.R. (1967), *Dictionary of Statistical Terms*, The International Statistical Institutes, New York.

[39] Kennedy, P. (1981). "Ballentine: A Graphical Aid for Economics." *Australian Economics Papers*, **20**, 414-416.

[40] Koenker, R.W.(1982), "Robust Methods in Econometrics," *Econometric Reviews*, **1**, 213-290.

[41] Lawrance, A.J.(1995), "Detection Influence and Masking in Regression," *Journal of the Royal Statistical Society* ,Series B, **57**, 181-189.

[42] Marquardt, D.W.(1970), "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," *Technometrics*, **12**, 591-612.

[43] Miller, R.G. (1974), "An Unbalanced Jackknife," *The Annals of Statistics*, **2**, 880-891.

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

[45] Newman, T.G. and Odell, P.L. (1971), *The Generation of Random Variates*, Charles Griffin, London.

[46] Pearson, E.S. and Chandra Sekar, C. (1936), "The Efficiency of Statistical Tools and a Criterion of Rejection of Outlying Observations," *Biometrika*, **28**, 308-320.

[47] Pearson, E.S. and Please, N.W. (1975), "Relation between the Shape of Population Distribution and the Robustness of Four Simple Statistical Tests," *Biometrika*, **62**, 223-241.

[48] Pearson, K. (1905), "On the General Theory of Skew Correlation and Non-linear Regression," *Biometrika*, **4**, 171-212.

[49] Pearson, K. and Lee A. (1903), "On the Laws of Inheritance," *Biometrika*, **2**, 357-462.

[50] Peña, D., and Yohai, V.J. (1995), "The Detection of Influential Subsets in Linear Regression by Using an Influence Matrix," *Journal of the Royal Statistical Society*, Series B, **57**, 145-156.

[51] Rousseeuw, P.J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, **79**, 871-880.

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

[53] Ryan, T. P. (1997), *Modern Regression Methods,* Wiely, New York.

[54] Sen, A. and Srivastava, M. (1990) *Regression Analysis: Theory, Methods and Applications*, Springer, New York.

[55] Silvey, S. D. (1969), "Multicollinearity and Imprecise Estimation," *Journal of the Royal Statistical Society, Series B*, **31**, 539-552.

[56] Staudte, R.G., and Sheather, S.J. (1990), *Robust Estimation and Test*, Wiley New York.

[57] Tsai, C., Cai, Z. & Wu, X. (1998). The Examination of Residual Plots. *Statistica Sinica* **8**, 445-465.

[58] Tukey, J.W. (1960), "A Survey of Sampling from Contamination Distributions," *Contribution to Probability and Statistics,* **1**, Olkin et. al. (Eds), 448-485.

[59] Velleman, P.F.,and Welsch, R.E.(1981), "Efficient Computing of Regression Diagnostics," *The American Statistician*, **35**, 234-242.

[60] Wetherill, G.B. (1986), *Regression Analysis with Applications*, Chapman and Hall, London.