2008

Robust Diagnostic Deletion Techniques in Linear and Logistic Regression

Nurunnabi, Abdul Awal Md.

University of Rajshahi

ROBUST DIAGNOSTIC DELETION TECHNIQUES IN LINEAR AND LOGISTIC REGRESSION

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Philosophy

By

Abdul Awal Md. Nurunnabi

Registration No. 41547, Roll No. 125 of Session July 2005

Department of Statistics

University of Rajshahi

Rajshahi, Bangladesh

March 2008

Dedication

To those who sacrificed their lives,
for
human civilization,
human rights,
peace on earth, and
against the war.

Abstract

Identification of unexpected observations is a topic of great attention in modern regression analysis. At the beginning statisticians differ but now they recognize robust regression and regression diagnostics are two complementary remedies to study unusual observations. We use both of them for identifying irregular observations at a time. We find out the group deletion diagnostic methods that show better performance for identifying influential observations in linear regression. These are based on robust regression and/or relevant diagnostic methods so that these are free from huge computational tasks and reliable in presence of masking and/or swamping because of prior suspect-group identification. We find a technique that performs well in case of large number and high-dimensional data sets. We have done a classification task of unusual observations in linear regression according to their nature of consequences on the analysis, and model building process. At the same time the method performs well for identifying influential observations. This method may be a good addition to the existing graphical literature. We have seen that our proposed procedures in linear regression are also effective to the logistic regression after some modification and development to the existing identification techniques in linear regression. Our further contribution is to propose two new identification techniques for influential observations in logistic regression. The new methods show efficient performance for the proper identification of unusual observations and thereby provide less misclassification error in the response variable for the binomial logistic regression. Summarizing all the above issues we can say that we have made contribution in three areas: identification of influential observations in linear regression, classification of unusual observations in linear regression, and identification of unusual observations in logistic regression.

Acknowledgements

"What soon grows old? Gratitude."

Aristotle

First and foremost I would like to thank my thesis supervisors, Professor Mohammed Nasser and Professor A. H. M. Rahmatullah Imon. This thesis would have not been possible without their spontaneous guidance and continuous support.

I owe to my teachers, friends and officials in the Department of Statistics, Rajshahi University, especially Professor Samad Abedin who taught me regression analysis in my class.

My eternal gratitude goes to my parents who showed me the first sun light on earth and that put my education above their personal comfort.

I acknowledge the valuable time, helps and advice of Professor A. C. Atkinson of the London School of Economics, UK, Professor Randy Sitter of the University of Simon Fraser, Canada, Professor Ali S. Hadi of the Cornell University, USA, Professor M. Nurul Islam of Dhaka University, Bangladesh.

I intend to thank all of my teachers, my friends, my relatives, my colleagues, my students and all others who have helped me directly or indirectly in my study and research.

I must express gratitude to my beloved younger brother and sister, and S. Akter for their constant inspiration and love.

At last, but not least, I bow to almighty Allah for allowing me the joy of writing and completing this research in time.

Abdul Awal Md. Nurunnabi

Declaration

I hereby declare that the thesis titled "Robust Diagnostic Deletion Techniques in Linear and Logistic Regression" is the result of my own research work for the degree of Master of Philosophy in the Department of Statistics, University of Rajshahi, Bangladesh. I further declare that this thesis has not concurrently been submitted anywhere.

Abdul Awal Md. Nurunnabi
M. Phil. Fellow
Roll No. 125
Registration No. 41547
Session, July 2005
Department of Statistics
University of Rajshahi
Rajshahi, Bangladesh

Certificate

This is to certify that the thesis titled "Robust Diagnostic Deletion Techniques in Linear and Logistic Regression" is a research work, carried out by Abdul Awal Md. Nurunnabi, for the degree of Master of Philosophy, in the Department of Statistics, University of Rajshahi, Bangladesh. This thesis has not been submitted for any degree or award anywhere.

M. Nasser
22/03/08
On behalf of

Mohammed Nasser, Ph D
Supervisor &
Professor
Department of Statistics
University of Rajshahi
Rajshahi
Bangladesh

Professor
Department of Statistics
Rajshahi University

Contents

List of Tables

List of Figures

Chapter 1

"That which is statistic and repetitive is boring,
that which is dynamic and random is confusing, in between lies art."

John Locke

Introduction

1.1 Regression Analysis and Its Historical Background

Regression analysis is a statistical technique, most widely used in almost every field of research and application in multifactor data, which helps us to investigate and to fit an unknown model for quantifying relations among observed variables. "It is appealing because it provides a conceptually simple method for investigating functional relationships among variables," (Chatterjee and Hadi, 2006). The standard approach in regression analysis is to take data, fit a model, and then evaluating the fit using various statistics.

First regression-type problems were considered in the 18^{th} century to aid navigation with the use of Astronomy and used exclusively in Physical science under Galton (cousin of Darwin). Galton in 1870 studied the question of quantifying genetic inheritance (intelligence, weight of sweet peas, heights of fathers and sons). He observed that sons of tall (short) fathers tend to be tall (short) as their fathers. He termed the phenomenon "regression to the mean". Pearson in 1896 formulated the idea of correlation in its most complete form, and developed ordinary least squares (OLS) as a method of parameter estimation in regression.

1.2 Linear Regression

Let us have a set of n observations $(y_i, x_i); i = 1,2,...,n$ of $(p+1)$ dimensional random vector (y, x) and want to quantify relation between y and x. To serve the purpose in regression analysis, the classical model assumes a relation of the scalar type

$$y_i = \beta_0 + x_{i_2}\beta_1 + ... + x_{i_k}\beta_p + \varepsilon_i, \quad i = 1,2,...,n \tag{1.1}$$

The relationship between Y (response) and $X_1, X_2, ..., X_p$ (predictors) is formulated as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon . \tag{1.2}$$

We can rewrite the system of equation by matrix notation as,

$$Y = X\beta + \varepsilon , \tag{1.3}$$

where $Y_{(n\times 1)}$ is the vector of response, $X_{n\times(p+1)}$ is the design matrix , $\beta_{(p+1)\times 1}$ is the parameter vector (regression coefficients) and $\varepsilon_{n\times 1}$ is error-vector. The adjective linear has a dual role, one is the relationship between response and the predictor(s) is linear and other is the model is linear in the parameters. It is assumed that any set of fixed values of $X_1, X_2, ..., X_p$ that fall within the range of the data, the linear equation (1.2) provides an acceptable approximation of the true relationship between Y and the X's. In particular, ε contains no systematic information for determining Y that is not already captured by the X's.

1.3 Logistic Regression

Logistic regression, a type of generalized regression, has been widely used since last two decades or so. From its original acceptance in epidemiological research, the method is now commonly used in fields including biomedical research, business and finance, criminology, engineering, wildlife biology, biometrics, data mining etc. Classically, logistic regression model is fitted to data obtained under experimental conditions; current use of these methods includes the analyses of data obtained in observational studies (Pregibon, 1981). Logistic regression is useful in situations for which we want to study

about the prediction of presence or absence of a specific characteristics or outcome based on values of a set of predictor variables. It is similar to a linear regression model but it is more suitable when the dependent variable is dichotomous, not continuous type. Logistic regression may use one of three types of categorical response variables: binary, ordinal, or nominal. Binary logistic regression is a form of regression, which is used when the dependent is a dichotomy and the independents are of any type. Multinomial logistic regression exists to handle the case of dependents with more than two classes. When multiple classes of the dependent variable can be ranked, then ordinal logistic regression is preferred to multinomial logistic regression.

1.4 Robustness and Robust Statistics

Box (1953) first introduced the technical term 'robust' and the subject matter acquired recognition as a legitimate topic for investigation only in mid-sixties, mainly due to the pioneer works of Tukey (1960,1962), Huber (1964), and Hampel (1968). Exact mathematical theory and probably the growing general awareness of the need for robust procedures due to the work E. S. Pearson, G. E. P. Box, and J. W. Tukey and others (see Hampel *et al.*, 1986; Huber, 1981; Maronna *et al.*, 2006; Nasser, 2000) have been brought robust statistics at this present stage.

Many assumptions commonly made in statistics (such as normality, linearity, independence) are at most approximations to reality. A minor error in the mathematical model should cause only a small error in the final conclusions. But this does not always hold, some of the most common statistical procedures are excessively sensitive to seemingly minor deviations from the assumptions. One reason is the occurrence of gross errors, such as copying or keypunch errors. They usually show up as outliers and are dangerous for many statistical procedures. Other reasons behind deviations from initialized model assumptions include the empirical characters of many models and the approximate characters of many theoretical models. The problem with the theories of classical parametric statistics is that they derive optimal procedures under exact

parametric models, but say nothing about their behaviors (stability) when the models are only approximately valid. In this regards, robust statistics try to study with 'optimality' and 'stability' both the mutually complementary characteristics in the same study. It is concerned with evaluating and improving the stability of estimation techniques when data points are deviated from assumptions. According to Davies and Gather (2004), the basic philosophy of robust statistics is to produce statistical procedures which are stable with respect to small changes in the data or model and even large changes should not cause a complete breakdown of the procedures. Hampel *et al.* (1986) mentioned, "Robust statistics, as a collection of related theories, is the statistics of approximate parametric models. In a broad informal sense, robust statistics is a body of knowledge, partly formalized into 'theories of robustness', relating to deviations from idealized assumptions in statistics. The theory of robustness is not just a superfluous mathematical decoration. It plays an essential role in organizing and reducing information about the behavior of statistical procedures to a manageable form".

1.5 Robust Regression

It is evident that least squares estimator (LSE) is extremely sensitive to atypical data and violations of its assumptions. Lack of stability of the LSE is not the only serious problem for estimating the parameters but also for the lack of normality assumptions on error terms we cannot test the reliability of the estimated parameters by using the common test procedures, we cannot check the model adequacy. Therefore, we depend on robust regression that possesses some stability in variance and bias under deviation from the regression model. A robust regression first wants to fit a regression to the majority of the data and then to discover the outliers as those points that possess large residuals from the robust output. Hence, the goal of robust regression is to safeguard against deviation from the assumptions of the classical least squares. Most popular robust regression techniques are LMS (least median squares) regression, LTS (least trimmed square) regression, reweighted least squares regression (RLS), M and MM- estimators.

1.6 Diagnostics

Diagnostics is another statistical approach that has been developing since sixties of the last century with robust statistics side by side to handle departures from strict parametric models. First use of this technique traced back about mid-nineteenth century (Barnett and Lewis, 1995). Diagnostics have taken traditionally a somewhat different view from robust statistics. Rather than modifying the fitting method, diagnostics condition on the fit using standard methods to attempt to diagnose incorrect assumptions, allowing the analyst to modify them and refit under the new set of assumptions, (Stahel and Weisberg, 1991).

1.7 Regression Diagnostics

"Regression diagnostics are techniques for exploring problems that compromise a regression analysis and for determining whether certain assumptions appear reasonable" (Fox, 1993). Field diagnostics is a combination of graphical and numerical tools. It is designed to detect and delete the outliers first and then to fit the good data by classical (least squares) methods. The basic building blocks of regression diagnostics are residuals, leverage values, vector of forecasts and vector of estimated parameters. The usual regression outputs clearly do not tell the whole story about the cause and/or effect of deviations from the assumptions of the model building process. Regression diagnostic can serve as the identification purpose of the deviations from the assumptions. So that basic need of regression diagnostics is to identify the unusual observations of a data set. Most popular diagnostic methods are Cook's distance, DFFITS, DFBETA, Atkinson's statistics, *etc.*

1.8 Interrelation between Robust Regression and Regression Diagnostics

There seems to be much confusion, even among workers in the two fields, about what robust and diagnostic methods are supposed to do. According to Huber(1991), "Robustness and diagnostics are complementary approaches to the analysis of data, and any one of the two is not good enough." Rousseeuw and Leroy (1987) mentioned, the purpose of robustness is to safeguard against deviations from the assumption; the purpose of diagnostics is to find and identify deviation from the assumptions. It means that each views the some problem from the opposite sites, and counting the metaphor, the more opaque the problem is, and the more important it is to view the problem from the all sides. They also mentioned, when using diagnostic tools, one first tries to delete the outliers and then to fit the 'good' data by least squares, whereas a robust analysis first wants to fit a majority of the data and then to discover the outliers as those points that possess large residuals from the robust solution. In robust regression, new procedures have been developed from theoretical considerations. Regression diagnostics, on the other hand, have been designed to supplement standard methodology with both graphical and non-graphical procedures.

1.9 Limitations of Existing Diagnostic Methods

Diagnostic methods have some limitations as follows:

1. Most of the popular diagnostic methods are based on the methods, which measures the sensitivity of a single observation at a time. Single diagnostic methods are failed to detect the unusual observations in presence of masking and/or swamping phenomena.
2. In case of group deletion diagnostics, it is a cumbersome and sometimes impossible to identify suspect group of unusual observations because of the sample size.
3. Diagnostic techniques those do not consider robustness can make non-robust decision.
4. It is hard to derive exact/asymptotic distribution of estimates obtained after classical diagnostic methods.

1.10 Why Robustness and Robust Regression

Robustness does not deal well with large deviation from model. First, most robust procedures geared towards large deviations lose uncomfortably much efficiency at "good" data sets. Second, for "bad" data set, any automated procedure may produce parameter estimates whose values are just as irrelevant for model interpretation as those of its non-robust siblings. On the other hand, important and large deviation can easily be missed by non-robust diagnostics, robust diagnostics are needed.

Robust regression estimates the regression parameter using the procedures that are insensitive to outliers (unusual observation). In regression diagnostics outlier (unusual observation) diagnostic is the main purpose and generally use the procedures which are sensitive to the outliers (unusual observations). Robust regression can help us to reduce the huge computational complexity and can meet as a first criterion of the identification techniques.

1.11 Classification of Unusual Observations

Observations are unusual in the sense that they are exceptional, they have extra role on model building process, or they may come from other population(s) and do not follow the pattern of the majority of the data. The presence of unusual observations could make huge interactive problems in inference. Because they can unduly influence the results of the analysis, and their presence may be a signal that a regression model fails to capture important characteristics of the data.

1.11.1 Outliers, Leverage Points and Influential Observations

In regression analysis, generally we can categorize unusual observations into three: outliers, high leverage points and influential observations. According to Hawkins (1980), an outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. In the scale parameter context, by an outlier we mean an observation that is so much larger than the bulk of the

observations that it stands out, and that there is doubt about it being from the proposed model. A rule suggests: an observation as outlying if it is more than three times the inter-quartile range from the median. (Staudte and Sheather, 1990). An observation that is apart from the bulk of the data is treated as outlier. In regression, outliers can be deviated into three ways (i) the deviation in the space of explanatory variable(s), deviated points in x-direction called leverage points (leverage points are two types: a) *good leverage*, if (x_i, y_i) does fit the linear relation, it improves the precession of the regression coefficient, and b) *bad leverage*, cases for which an x_i is far away from the bulk of the x_i , do not fit linear relationship (ii) the change in the direction of response (Y) variable (outlier in Y-direction but not a leverage point is called vertical outlier) generally measured by absolute magnitude of standardized/Studentized residual of the observation, (iii) the other is change in both the directions (x_i; direction of the explanatory variables and y_i; direction of the response variable). Influential observation "is one which either individual or together with several other observations has a demonstrably larger impact on the calculated values of various estimates than is the case for most of the other observations" (Belsley *et al.* 1980). It is to be noted that an outlier or a leverage point is not necessarily an influential observation and the converse is also true, that is an influential observation may not be an outlier or a leverage point.

1.11.2 Classification Techniques and Their Limitations

Generally classification tasks of unusual observations are performed by identification of outliers, high-leverage points and influential observations separately. Most of the methods of identification are distance based and some well-known methods are Mahalanobis distance, robust distance (RD), minimum covariance determinant (MCD), and minimum volume ellipsoid (MVE). Most of the popular diagnostic techniques for identifying outliers and high-leverage points focusing on both of them separately and they do not have any combined visual perception power with respect to their influence on analysis and decision making process.

1.12 Objectives of the Study

The identification of unusual observations is an issue of a great attention in regression analysis. Classical diagnostics based on LS estimates often fail to reveal unusual observations. Robust regression and regression diagnostics are two complementary approaches (remedy) to deal with unusual observations in regression analysis. Regression diagnostic methods perform their tasks by identifying unusual observations and studying the sensitivity of the statistics engaged in identifying purpose.

Our main objectives in this study are:

i. To study the existing diagnostic methods for influential observations in linear and logistic regression.

ii. To expound the necessity of group deletion technique and use of robust regression for identification task.

iii. To develop and propose some new identification techniques for influential observations in presence of masking and swamping phenomena in linear and logistic regression.

iv. To develop diagnostic techniques for high dimensional and large data set.

v. To study the basic properties of leverage and residual matrices and the regression diagnostics measures.

1.13 Thesis Overview

We present a short review of ideas about regression, linear regression, logistic regression, robustness, robust regression, regression diagnostics, types of unusual observations and objectives of my thesis in Introduction, Chapter 1. The rest of the thesis is organized as follows.

Chapter 2 contains some preliminary mathematics for regression diagnostics and data source. It provides some matrix algebra concepts, (such as inverse, generalized inverse, partitioned matrix, eigen-structure, *etc.*) some fundamental properties relevant with the

leverage and residual matrices, deletion diagnostic algebra that are urgently needed for the calculation purpose, and existing methods and their interactive relations side by side.

Chapter 3 is a literature review of the two: robust Regression and regression diagnostics. It deals with some popular approaches, fundamental ideas in both of them and interrelations between them.

Chapter 4 proposes two new measures based on deletion idea for identifying influential observations in linear regression. At the same time chapter provides some comparisons showing better performance with existing popular methods through some well-referred data sets.

Chapter 5 possesses a five-fold plotting technique on a potential-residual (P-R) plot with robust distance that can separate unusual observations from the regular and the technique classify unusual observations into: outliers, high-leverage points and influential observations. Several demonstrations are also performed.

Chapter 6 presents different diagnostic aspects in logistic regression. It introduces two diagnostic measures for the identification of multiple influential observations in binomial logistic regression.

Chapter 7 makes the conclusions and some indications of further research.

Appendix A provides the data sets that are used in this thesis.

Appendix B attaches abstracts of the articles that are already published, accepted or submitted for the publication. Those are related to our research work.

Chapter 2

"True genius lies in the capacity for evaluation of [...] conflicting information."

Winston Churchill

Preliminary Mathematics for Diagnostics and Data Source

Regression diagnostics by deletion of one or a group of observations (cases) are mostly depending on some matrix algebra. Inverse matrix, generalized inverse, different partitioned matrices and eigen-structure of a matrix are the basics of deletion diagnostics. To study the deletion diagnostic algebra we have studied necessary diagnostic algebra with their fundamental properties. The algebra of least squares technique is presented as a cornerstone of regression diagnostic method. Prediction (leverage) and residual matrices are the two basic building blocks of deletion diagnostics. Several new, interesting and useful properties of leverage and residual matrices for deletion diagnostics are developed and presented in this chapter. Algebra of most popular diagnostic measures is presented side by side for the completion of the thesis. Data sources are also given here.

2.1 Least Squares Algebra

To fit a regression model we estimate the parameters, and the most well known method is least squares (LS) method. From the time of its invention to last quarter of the last century it was the cornerstone of regression analysis for its mathematical beauty and computational simplicity. The underlying principle of this method is to estimate the

unknown parameters of a regression model in such a way that the total squared deviations of the observed and fitted response would be minimized. Most of the diagnostic methods are based on LS or originated from the idea of LS. In favor of our study here we consider the algebra of least squares.

2.1.1 Least Squares Estimation (LSE)

It is customary that the classical linear model can be defined in matrix notation as

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

where Y is an $n \times 1$ vector of response (continuous) variable, X is an $n \times k(n > k = p+1)$ matrix formed by explanatory variables with a constant, β is a $k \times 1$ unknown vector of parameters with a constant, and ε is the vector of identically and independently distributed (*i.i.d*) random error terms. The method of least squares (LS), minimizes the error sum of squares or equivalently, finds the vector of LS estimatorsβ, which minimizes

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta). \tag{2.2}$$

The least squares estimators must satisfy

$$\left.\frac{\delta S}{\delta \beta}\right|_{\beta} = -2X^T Y + 2X^T X\hat{\beta} = 0, \tag{2.3}$$

which simplifies to

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

Equation (2.4) is the least squares normal equation and is identical to (2.2). To solve the normal equation, premultiply both sides of (2.4) by the inverse of $X^T X$. Thus the least squares estimate of β is

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

The fitted regression model corresponding to the level of the regressor variables is,

$$\hat{Y} = X^T \hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j. \tag{2.6}$$

The corresponding residual vector

$$\hat{\varepsilon} = r = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X^T X)^{-1} X^T Y \quad (2.7)$$

$$= Y - HY = (I - H)Y \quad (2.8)$$

$$= (I - H)(X\beta + \varepsilon) \quad (2.9)$$

$$= (I - H)\varepsilon \ , \quad (2.10)$$

where $H = X(X^T X)^{-1} X^T$ is the leverage or prediction or hat matrix. In scalar form, i-th residual is

$$r_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j ; \quad i = 1,2,...,n . \quad (2.11)$$

Clearly, if the h_{ij} are sufficiently small, r_i will serve as a reasonable alternative of ε_i.

2.2 Essential Matrix Algebra for Deletion Diagnostics

To serve the purpose of deletion diagnostic algebra we attach some fundamental concepts of matrix algebra that are urgently needed in almost in all respect of computations of this thesis. We try to arrange the ideas according to their technical necessity.

2.2.1 Rank and Inverse of a Matrix

Rank and inverse are two most useful and important characteristics of a matrix and play a vital role throughout all aspects of matrix and also for diagnostic algebra.

The rank of a matrix is the number of linearly independent (LIN) rows (columns) in the matrix. The rank of A is denoted by r_A or $r(A)$. The following properties and consequences of rank are important.

i. r_A is a positive integer, except that r_0 is defined as $r_0 = 0$.

ii. $r(A_{p\times q}) \le p \; and \; \le q$: the rank of a matrix equals or less than the smaller of its number of rows or columns.

iii. $r(A_{n\times n}) \le n$: a square matrix has rank not exceeding its order.

iv. When $r_A = r \neq 0$ there is at least one square sub matrix of A having order r that is nonsingular.

v. $r(A_{n\times n}) = n$ then by (iv) A is nonsingular, *i.e.*, A-inverse exists.

vi. $r(A_{n\times n}) < n$ then A is singular and A-inverse does not exist.

vii. $r(A_{p\times q}) = p < q$, A is said to have full row rank, or to be of full row rank. Its rank equals its number of rows.

viii. $r(A_{p\times q}) = q < p$, A is said to have full column rank. Its rank equals its number of columns.

ix. $r(A_{n\times n}) = n$, A is said to have full rank. Its rank equals its order, it is nonsingular, its inverse exists.

The concept of a matrix inverse has been established in the context of solving simultaneous linear equations. We need this to estimate the parameters or to fit the values of Y and for finding the values of residuals or leverage matrices.

The inverse of a square matrix A is a matrix whose product with A is the identity matrix. It is denoted by the symbol A^{-1} and is read as "the inverse of A" or as "A- inverse", *i.e.*, $AA^{-1} = I$. Now the basic question: "Is there a matrix L whose product with the matrix A is I?" There are three answers to this, depending on the characteristics of A:

i) In some cases L exists and is unique for a given A.

ii) Some times numerous L exists *i.e.,* L is not unique, and

iii) In some instances L does not exist at all.

Generalized Inverse, a generalized system of matrix inverse. A $n\times k$ rectangular (singular) matrix A possess either a left inverse or a right inverse when $r(A) = \min(n, k)$, when a matrix is not of full rank, neither a left nor a right inverse exists, and as a result we cannot always use a matrix inverse to solve equations. The notion of a generalized inverse can then be easily extended to inconsistent linear systems. It plays a major role in regression estimation and in leverage matrix.

Moore-Penrose Inverse: Given any matrix A, there is a unique matrix M such that

(i) AMA=A (iii) AM is symmetric

(ii) MAM=M (iv) MA is symmetric.

This result (above four) is developed in Penrose (1955), on foundations laid by Moore (1920). The Penrose paper established not only the existence of M but also its uniqueness for a given A. One-way of writing M is based on the factoring of $A_{p\times q}$ as A=KL, where K and L have full column and row rank respectively, equal to $r(A)$. Then M of the above four conditions is

$$M = L^T (K^T A L^T)^{-1} K^T . \tag{2.12}$$

The matrix M defined by the four Penrose conditions is unique for a given A. But there are many matrices G which satisfy just the first Penrose condition:

$$AGA=A. \tag{2.13}$$

Any matrix G satisfying (2.13) is called a generalized inverse of A; when A is $p\times q$ then G is $q\times p$.

Inverse of a Partitioned Matrix

In case of group deletion diagnostics, we need to partition the design or the augmented matrices and these are partitioned as of suspect cases and regular cases but it is not always possible to exist inverse for all the partitioned matrices. We need the help of the following fundamental formulae/properties of the partitioned inverse.

Theorem 2.2.1 (Dhrymes, 1984)

Let A be a square nonsingular matrix of order m, and partition as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

such that A_{ii}, i =1 and 2, are nonsingular matrices of order m_i, i =1 and 2, respectively $(m_1 + m_2 = m)$.Then

$$B = A^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \tag{2.14}$$

where

$$B_{11} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}, \quad B_{12} = -A_{11}^{-1}A_{12}\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1},$$
$$B_{21} = -A_{22}^{-1}A_{21}\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}, \quad B_{22} = \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}.$$

Theorem 2.2.2 (Chatterejee and Hadi, 1988)

Let A be a matrix partition as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

(a) If A and A_{11} are nonsingular, then

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}MA_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}M \\ -MA_{21}A_{11}^{-1} & M \end{bmatrix}, \tag{2.15}$$

where $M = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$.

(b) If A and A_{22} are nonsingular, then

$$A^{-1} = \begin{bmatrix} N & -NA_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{12}N & A_{22}^{-1} + A_{22}^{-1}A_{21}NA_{12}A_{22}^{-1} \end{bmatrix}, \tag{2.16}$$

where $N = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$.

2.2.2 Eigen Values and Eigen Vectors

It is known that eigenstructure of a matrix X can change significantly when a row of the design matrix is added to or omitted from X. we can study the influence of the i-th row of X on the eigenstructure of X in general and on its condition number and collinearity indices. Let A be a square matrix of dimension $k \times k$ and λ be an eigen value of A. If $x_{k\times 1}$ is a nonzero vector such that

$$Ax = \lambda x . \tag{2.17}$$

Then x is said to be an eigenvector (characteristic vector) of the matrix A associated with the eigen value λ.

Another point of view; Let A be a $k \times k$ square matrix and I be the $k \times k$ identity matrix. Then the scalars $\lambda_1, \lambda_2, ..., \lambda_k$ satisfying the polynomial equation $|A - \lambda I| = 0$ are called the *eigen values* of a matrix A.

The equation

$$|A - \lambda I| = 0 \tag{2.18}$$

(as a function of λ) is called the characteristic equation.

2.2.3 Grammian Matrix and Its Properties

If A be $n \times m$ real matrix then the matrix $S = A^T A$ is called Grammian matrix of A. If A is $m \times n$ then $S = A^T A$ is a symmetric n-rowed matrix. It has the following important properties.

i. Every positive definite or positive semidefinite matrix can be represented as a Grammian matrix.

ii. The Grammian matrix $A^T A$ is always positive definite or positive semidefinite, and it depends on the rank of A or less than the number of its column accordingly.

iii. The rank of the matrix $A^T A$ is same as that matrix A and AA^T. That is if $r(A) = r \quad then \quad r(A^T A) = r(AA^T) = r$

iv. If $(AA^T) = 0 then \quad A = 0.$

2.2.4 Idempotent Matrices and Projections

Idempotent matrix, A square matrix A is said to be idempotent when $A = A^2$. For any idempotent matrix we have $A = A^r$ for any integer $r>0$.

Theorem 2.2.3 (Basilevsky, 1983)

An idempotent matrix A is always singular, except for the unit matrix I.

Theorem 2.2.4 (Basilevsky, 1983)

Let A and B be idempotent matrices. Then we have the following:

i. $A + B$ is idempotent only when $AB=BA=0$.

ii. $C=AB$ is idempotent only when $AB=BA$.

iii. $I-A$ is idempotent.

Property2.2.1 (Chatterjee and Hadi 1988)

Let A be an $n \times n$ idempotent matrix. Then we have the following:

(i) The eigen values of A are equal to 0 or 1.

(ii) The trace and rank of A are equal.

Projection and Projection Matrix

Consider an arbitrary vector $Y \in V$, where the vector space $V = V_1 \oplus V_2$ is decomposed as a direct sum ($\oplus$) of V_1 and V_2. Also let $Y = Y_1 + Y_2$, where $Y_1 \in V_1$ *and* $Y_2 \in V_2$. Then the transformation $PY=Y_1$ is called the projection of vector Y onto the vector space V_1 along the vector space V_2 if and only if $PY_1=Y_1$, P is defined as a projection matrix, that projects vector Y onto a subspace. P has the following:

i. P is associated with a linear transformation.

ii. P is a projection matrix if and only if P is an idempotent matrix.

Matrices of the form $(X^T X)^{-1} X^T$ occur often in linear regression models. This is the matrix that transforms the response vector Y into the least squares estimates of β in the linear model $Y = X\beta + \varepsilon$, for example. It is a factor of the projection matrix of Y onto $\hat{Y}$, that is, $X(X^T X)^{-1} X^T$, the "hat" matrix. We recall that a matrix is a projection matrix if and only if it is symmetric and idempotent (that means a projection matrix is necessarily either the identity or it is singular).

2.2.5 Decompositions

Decompositions provide a numerically stable way to solve a system of linear equations, as shown already in, and to invert a matrix. Additionally, they provide an important tool for analyzing the numerical stability of a system. Decompositions allow us to transform a general system of linear equations to a system with an upper triangular, a diagonal, or a lower triangular coefficient matrix. Some of most frequently used decompositions are the Cholesky, QR, LU, and SVD decompositions. Now we define the following two to serve our purpose.

The Spectral Decomposition, Let A be a $k \times k$ symmetric matrix. Then A can be express in terms of its k eigenvalue-eigenvector pairs (λ_i, e_i) as

$$A = \sum_{i=1}^{k} \lambda_i e_i e_i' \quad . \tag{2.19}$$

Singular-Value Decomposition, Let A be an $m \times k$ matrix of real numbers. Then there exist a $m \times m$ orthogonal matrix U and a $k \times k$ orthogonal matrix V such that

$$A = UAV^T, \tag{2.20}$$

where the $m \times k$ matrix A has *(i,i)* entry $\lambda_i \geq 0$ for $i=1,2,...,min(m,k)$ and other entries are zero. The positive constant λ_i is called the singular value of A.

2.2.6 Fundamental Deletion Formula

Following results are useful for finding the inverse of leverage matrices when one or a group of observations are added to or omitted from others.

(i) Let A be a nonsingular matrix and U and V two column vectors, then

$$\left(A - UV^T\right)^{-1} = A^{-1} + A^{-1}U\left(I - V^T A^{-1}U\right)^{-1} V^T A^{-1} \tag{2.21}$$

Proof:

$$\begin{aligned}
&\left(A - UV^T\right)\left\{A^{-1} + A^{-1}U\left(I - V^T A^{-1}U\right)^{-1} V^T A^{-1}\right\} \\
&= I + U\left(I - V^T A^{-1}U\right)^{-1} V^T A^{-1} - UV^T A^{-1} - UV^T A^{-1}U\left(I - V^T A^{-1}U\right)^{-1} V^T A^{-1} \\
&= I - UV^T A^{-1} + U\left(I - V^T A^{-1}U\right)\left(I - V^T A^{-1}U\right)^{-1} V^T A^{-1} \\
&= I - UV^T A^{-1} + UV^T A^{-1} = I .
\end{aligned}$$

(ii) $(A + UV^T)^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1}U)^{-1} V^T A^{-1}$; (Rao, 1973) (2.22)

(iii) Let A and D be nonsingular matrices of orders k and m respectively, B be $k \times m$, and C be $k \times m$. Then, provided that the inverses exist,

$$\left(A + BDC^T\right)^{-1} = \left\{A^{-1} - A^{-1}B\left(D^{-1} + C^T A^{-1}B\right)^{-1} C^T A^{-1}\right\} \tag{2.23}$$

Proof:

$$\begin{aligned}
&\left(A + BDC^T\right)\left\{A^{-1} - A^{-1}B\left(D^{-1} + C^T A^{-1}B\right)^{-1} C^T A^{-1}\right\} \\
&= I - B\left(D^{-1} + C^T A^{-1}B\right)^{-1} C^T A^{-1} + BDC^T A^{-1} - BDC^T A^{-1}B\left(D^{-1} + C^T A^{-1}B\right)^{-1} C^T A^{-1} \\
&= I + BDC^T A^{-1} - B\left(I + DC^T A^{-1}B\right)\left(D^{-1} + C^T A^{-1}B\right)^{-1} C^T A^{-1} \\
&= I + BDC^T A^{-1} - BD\left(D^{-1} + C^T A^{-1}B\right)\left(D^{-1} + C^T A^{-1}B\right)^{-1} C^T A^{-1}
\end{aligned}$$

$= I + BDC^T A^{-1} - BDC^T A^{-1}$

$= I$.

(Note, where the inverse of $(A + BDC^T)$ does not exist is found by taking A=I, B=X, $C^T = X^T$, *and* $D = -(X^T X)^{-1}$. *This yields* $(A + BDC^T) = (I - H)$, which is singular, but this is also computable when we take the advantage of generalized inverse.)

(iv) Consider the $k \times k$ matrix $X^T X$ and let x_i^T be the i-th row of X. then

$$\left(X^T X - x_i x_i^T\right)^{-1} = \left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} x_i x_i^T \left(X^T X\right)^{-1}}{1 - x_i^T \left(X^T X\right)^{-1} x_i} \tag{2.24}$$

Proof:

Multiplying right-hand side by $\left(X^T X - x_i x_i^T\right)$,

$$\left[\left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} x_i x_i^T \left(X^T X\right)^{-1}}{1 - x_i^T \left(X^T X\right)^{-1} x_i}\right]\left(X^T X - x_i x_i^T\right)$$

$$= I + \frac{\left(X^T X\right)^{-1} x_i x_i^T}{1 - x_i^T \left(X^T X\right)^{-1} x_i} - \left(X^T X\right)^{-1} x_i x_i^T - \frac{\left(X^T X\right)^{-1} x_i x_i^T}{1 - x_i^T \left(X^T X\right)^{-1} x_i} \left(X^T X\right)^{-1} x_i x_i^T$$

$$= I + \frac{\left(X^T X\right)^{-1} x_i x_i^T - \left(X^T X\right)^{-1} x_i x_i^T \left(1 - x_i^T \left(X^T X\right)^{-1} x_i\right) - \left(X^T X\right)^{-1} x_i \left| x_i^T \left(X^T X\right)^{-1} x_i \right| x_i^T}{1 - x_i^T \left(X^T X\right)^{-1} x_i}$$

$= I$.

2.3 Properties of Leverage and Residual Matrix

Chatterjee and Hadi (1986) call H as the prediction matrix because it is the transformation matrix that, when applied to Y, produces the predicted values. Similarly, *(I-H)* is the residual matrix, because applying it to Y produces the ordinary residuals. We discuss a comprehensive account of their properties.

Property 2.3.1 (Chatterjee and Hadi, 1988; Atkinson and Riani, 2000)

(a) Leverage matrix, *H and* (b) Residual matrix *(I-H)* are

(i) Symmetric matrix (ii) Idempotent matrix.

Proof:

We have $$H = X(X^T X)^{-1} X^T$$

(a) (i) $$H^T = (X(X^T X)^{-1} X^T)^T = X(X^T X)^{-1} X^T = H.$$

(ii) $$H^2 = HH = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = H.$$

(b) (i) $$(I - H)^T = [I - X(X^T X)^{-1} X^T]^T = I - X(X^T X)^{-1} X^T = I - H.$$

(ii) $$(I - H)^2 = (I - H)(I - H) = I - 2H + H^2 = I - 2H + H = I - H.$$

Property 2.3.2 (Chatterjee and Hadi, 1988)
Let X be $n \times k$, then

(a) $trace(H) = rank(H) = k$, (b) $trace(I - H) = n - k$ and, (c) $\sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^2 = k$

Proof:

(a) $$\begin{aligned} tr(H) &= tr\left\{X\left(X^T X\right)^{-1} X^T\right\} \\ &= tr\left\{\left(X^T X\right)^{-1} X^T X\right\} \\ &= tr(I_k) \\ &= k. \end{aligned}$$

(b) $$\begin{aligned} &tr(I - H) \\ &= tr(I) - tr(H) = n - k. \end{aligned}$$

(c) Since H is idempotent,

$$H = H^2$$

$$\Rightarrow h_{ii} = \sum_{j=1}^{n} h_{ij}^2$$

$$\Rightarrow \sum_{i} h_{ii} = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^2 = tr(H) = k. \qquad (2.25)$$

Property 2.3.3 (Chatterjee and Hadi, 1988)

For $i = 1,2,...,n$ *and* $j = 1,2,...,n$ we have

(a) $0 \le h_{ii} \le 1$ for all i and, (b) $-0.5 \le h_{ij} \le 0.5$ for all $j \ne i$.

Proof:

(a) By property 2.3.2 the i-th diagonal elements of H can be written as

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^{\ 2} = h_{ii}^{\ 2} + \sum_{j \neq i} h_{ij}^{\ 2}$$
$$i.e., \quad h_{ii}^{\ 2} \leq h_{ii} \tag{2.26}$$

from which it follows that $0 \leq h_{ii} \leq 1$, for all i.

(b) Identity (2.26) can also be expressed as

$$h_{ii} = h_{ii}^{\ 2} + h_{ij}^{\ 2} + \sum_{r \neq i,j} h_{ir}^{\ 2}$$

from which it follows that $h_{ij}^{\ 2} \leq h_{ii}(1 - h_{ii})$ and since $0 \leq h_{ii} \leq 1$ hence $h_{ij}^{\ 2}$ will be the largest when $h_{ii} = 0.5$, and thus $-0.5 \leq h_{ij} \leq 0.5$.

Property 2.3.4 (Chatterjee and Hadi, 1988)

Let $X = (X_1 : X_2)$ 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. Let $H_1 = X_1(X_1^{\ T}X_1)^{-1}X_1^{\ T}$ be the prediction matrix for X_1, and $W = (I - H_1)X_2$ be the projection of X_2 onto the orthogonal component of X_1. Finally, let $H_2 = W(W^{\ T}W)^{-1}W^{\ T}$ be the prediction matrix for W. Then H can be expressed as

$$X(X^TX)^{-1}X^T = X_1(X_1^{\ T}X_1)^{-1}X_1^{\ T} + (I - H_1)X_2\{X_2^{\ T}(I - H_1)X_2\}^{-1}X_2^{\ T}(I - H_1)$$

or
$$H = H_1 + H_2. \tag{2.27}$$

Proof:

$$H = X(X^TX)^{-1}X^T$$
$$= (X_1 \quad X_2)\begin{bmatrix} X_1^{\ T}X_1 & X_1^{\ T}X_2 \\ X_2^{\ T}X_1 & X_2^{\ T}X_2 \end{bmatrix}^{-1}\begin{bmatrix} X_1^{\ T} \\ X_2^{\ T} \end{bmatrix}$$

By using the form of inverse of a partitioned matrix (eq. 2.15),

$$(X^TX)^{-1} = \begin{bmatrix} (X_1^{\ T}X_1)^{-1} + (X_1^{\ T}X_1)^{-1}X_1^{\ T}X_2MX_2^{\ T}X_1(X_1^{\ T}X_1)^{-1} & -(X_1^{\ T}X_1)^{-1}X_1^{\ T}X_2M \\ -MX_2^{\ T}X_1(X_1^{\ T}X_1)^{-1} & M \end{bmatrix},$$

where
$$M = (X_2^{\ T}X_2 - X_2^{\ T}X_1(X_1^{\ T}X_1)^{-1}X_1^{\ T}X_2)^{-1}$$

$$= \{X_2^T (I - X_1 (X_1^T X_1)^{-1} X_1^T) X_2\}^{-1}$$

$$= \{X_2^T (I - H_1) X_2\}^{-1}.$$

Hence, $H =$

$$\left[X_1(X_1^TX_1)^{-1}+X_1(X_1^TX_1)^{-1}X_1^TX_2MX_2^TX_1(X_1^TX_1)^{-1}-X_2MX_2^TX_1(X_1^TX_1)^{-1} \quad -X_1(X_1^TX_1)^{-1}X_1^TX_2M+X_2M\right]\begin{bmatrix}X_1^T\\X_2^T\end{bmatrix}$$

$$= (X_1 (X_1^T X_1)^{-1} X_1^T + X_1 (X_1^T X_1)^{-1} X_1^T X_2 M X_2^T X_1 (X_1^T X_1)^{-1} X_1^T$$
$$- X_2 M X_2^T X_1 (X_1^T X_1)^{-1} X_1^T - X_1 (X_1^T X_1)^{-1} X_1^T X_2 M X_2^T + X_2 M X_2^T)$$

$$= H_1 + H_1 X_2 M X_2^T H_1 - H_1 X_2 M X_2^T - X_2 M X_2^T H_1 + X_2 M X_2^T, \text{ (1st part)}$$

$$= H_1 - H_1 X_2 M X_2^T + H_1 X_2 M X_2^T H_1 - X_2 M X_2^T H_1 + X_2 M X_2^T$$

$$= H_1 - H_1 X_2 M X_2^T (I - H_1) + X_2 M X_2^T (I - H_1)$$

$$= H_1 - (H_1 - I) X_2 M X_2^T (I - H_1)$$

$$= H_1 + (I - H_1) X_2 M X_2^T (I - H_1)$$

$$= H_1 + (I - H_1) X_2 \{X_2^T (I - H_1) X_2\}^{-1} X_2^T (I - H_1)$$

$$= H_1 + H_2.$$

Property 2.3.5 (Chatterjee and Hadi, 1988)

For $i = 1,2,...,n$ *and* $j = 1,2,...,n$

(a) If X contains a constant column, then

(i) $h_{ii} \geq n^{-1}$ for all i.

(ii) $H1 = 1$, where 1 is an n vector of ones.

(b) Suppose that the i-th row of X, x_i, occurs a times and that the negative of x_i occurs b times. Then $h_{ii} \leq (a+b)^{-1}$.

Proof:

(a) If X contains a constant column, define $X = (1 : X_2)$ where 1 is the n-vector of ones. From property 2.3.4, we have

$$H_1 = 1(1^T 1)^{-1} 1^T = n^{-1} 11^T;$$
$$W = (I - H_1)X_2 = (I - n^{-1}11^T)X_2 \equiv \tilde{X};$$
$$H_2 = \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T$$

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

Thus the prediction matrix H can be written

$$H = H_1 + H_2 = n^{-1}11^T + \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T.$$

Each of the diagonal elements of H_1 is equal to n^{-1} and since H_2 is a prediction matrix, by property 2.3.3 (a) ($0 \le h_{ii} \le 1$), its diagonal elements are nonnegative, hence $h_{ii} \ge n^{-1}$ for all i, since $\tilde{X}^T 1 = 0, H_2 1 = 0$ and thus $H1 = H_1 1 = 1$.

(b) Suppose that x_i occurs a times and $-x_i$ occurs b times.

Let $J = \{j : x_j = x_i \quad or\ x_j = -x_i \quad , j = 1,2,...,n \quad \}$ be the set of the row indices of these replicates. Since $h_{ij} = x_i^T (X^T X)^{-1} x_j$, then $h_{ii} = |h_{ij}|\ for \quad j \varepsilon J$ and by (eq.2.26) we can reexpress h_{ii} as

$$h_{ii} = \sum_{j \in J} h_{ij}^2 + \sum_{j \notin J} h_{ij}^2$$
$$= (a+b)h_{ii}^2 + \sum_{j \notin J} h_{ij}^2 \ge h_{ii}^2 (a+b),$$

from which it follows that $h_{ii} \le (a+b)^{-1}$.

Property 2.3.6 (Chatterjee and Hadi, 1988)

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

(a) If $h_{ii} = 0$ or 1, then $h_{ij} = 0$.

(b) $h_{ii} + \frac{r_i^2}{r^T r} \le 1$.

Proof:

(a) Since H is an idempotent matrix,

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

If $h_{ii} = 0$ then, $\quad 0 = 0 + \sum_{i \neq j} h_{ij}^2$

$$i.e., h_{ij} = 0;$$

if $h_{ii} = 1$, then

$$1 = 1 + \sum_{i \neq j} h_{ij}^2$$

$$i.e., h_{ij} = 0.$$

(b) Define $Z = (X: Y)$, $H_X = X\left(X^T X\right)^{-1} X^T$, and $H_Z = Z\left(Z^T Z\right)^{-1} Z^T$. We can write, (by virtue of eq.2.27)

$$H_Z = H_X + \frac{(I - H_X)YY^T(I - H_X)}{Y^T(I - H_X)Y}$$

$$= H_X + \frac{rr^T}{r^T r} \leq 1, \text{ since the diagonal elements of } H_z \leq 1$$

$$= h_{ii} + \frac{r_i^2}{r^T r} \leq 1 \quad . \tag{2.28}$$

Property 2.3.7 (Atkinson and Riani, 2000)

In simple regression h_{ii} will be large if x_i is far from the bulk of other points in the data.

Proof:

In the case of simple regression of Y on X (one variable) through the origin, we have

$$X = \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{pmatrix}; \quad X^T = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix},$$

so $$X^T X = \sum_{i=1}^{n} x_i^{\,2}\ ;\ \text{hence}\ H = X(X^T X)^{-1} X^T = XX^T \left(\sum_{i=1}^{n} x_i^{\,2} \right)^{-1}.$$

From which it follows that

$$h_{ii} = \frac{x_i^{\,T} x_i}{\sum_{i=1}^{n} x_i^{\,2}} = \frac{x_i^{\,2}}{\sum_{i=1}^{n} x_i^{\,2}}; \qquad i = 1,2,...,n.$$

If a constant term is included in the model, *i.e.*,

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix}.$$

In this case $X^T X$ *and* $(X^T X)^{-1}$ are respectively equal to

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^{\,2} \end{pmatrix} \textit{and} \quad (X^T X)^{-1} = \frac{1}{n\sum_{i=1}^{n} (x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^{\,2} & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}.$$

Then h_{ii} will be equal to

$$\frac{1}{n\sum_{i=1}^{n} (x_i - \bar{x})^2} \begin{pmatrix} 1 & x_i \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} x_i^{\,2} & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} 1 \\ x_i \end{pmatrix}.$$

$$= \frac{\sum_{i}^{n} x_i^{\,2} - 2x_i \sum_{i=1}^{n} x_i + n x_i^{\,2}}{n\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Adding and subtracting $n\bar{x}^2$ in the numerator we obtain:

$$h_{ii} = \frac{1}{n} + \frac{n\bar{x}^2 - 2x_i \sum_{i=1}^{n} x_i + nx_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

$$i.e., \quad h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.29}$$

Thus in simple regression h_{ii} will be large if x_i is far from the bulk of other points in the data.

Property 2.3.8 (Chatterjee and Hadi, 1988)

Let X be an $n \times k$ matrix of rank k. Then for fixed $n, h_{ii}, i = 1,2,\cdots,n$, is nondecreasing in k.

Proof: We have showed that $H = H_1 + H_2$, where H_1 and H_2 are prediction matrices. From the property 2.3.4 (eq.2.27), the diagonal elements of H_1 and H_2 are nonnegative. Therefore, for fixed $n, h_{ii}, i = 1,2,\cdots,n$ is non-decreasing in k.

2.4 Deletion Diagnostic Algebra

When an observation or a group of observations is deleted and the regression model refitted, the parameter estimates, leverage values, residuals, and residual sum of squares all will change. Statisticians observe the effects of deletion and measure the sensitivity of the diagnostic measures, so that the explicit deletion of individual observations and repeated refitting of the model are not necessary. The methods are collectively known as deletion diagnostics. Relevant algebra in these methods is the deletion diagnostic algebra.

2.4.1 Single Case Deletion

The following results have found by deletion of one case. Results show how the effect of deletion of one case impact on results from whole data set and on the estimates.

Property 2.4.1

When $h_{jk}^{(-i)}$ is the jk-th element of the matrix H excluding the i-th row of matrix X,

then (a) $$h_{jk}^{(-i)} = h_{jk} + \frac{h_{ji}h_{ik}}{1-h_{ii}}$$

(b) $$h_{jj}^{(-i)} = h_{jj} + \frac{h_{ij}^2}{1-h_{ii}}; \quad \text{when } j=k$$

(c) $$h_{ii}^{(-i)} = \frac{h_{ii}}{1-h_{ii}}; \quad \text{when } j=k=i$$

(d) For fixed k, $h_{jj}, j = 1,2,\cdots,n$ is nonincreasing in n.

Proof:

(a) We have

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

premultiplying both sides by $x_j{}^T$ and postmultiplying by x_k , we get

$$x_j{}^T (X_{(i)}{}^T X_{(i)})^{-1} x_k = x_j{}^T (X^T X)^{-1} x_k + \frac{x_j{}^T (X^T X)^{-1} x_i x_i{}^T (X^T X)^{-1} x_k}{1-h_{ii}}$$

$$\Rightarrow h_{jk}^{(-i)} = h_{jk} + \frac{h_{ji}h_{ik}}{1-h_{ii}}.$$

(b) If $j = k$

$$h_{jj}^{(-i} = x_j^T \left(X_{(i)}^T X_{(i)}\right)^{-1} x_j$$

$$= x_j^T \left(X^T X\right)^{-1} x_j + \frac{x_j^T \left(X^T X\right)^{-1} x_i x_i^T \left(X^T X\right)^{-1} x_j}{1-h_{ii}}$$

$$= h_{jj} + \frac{h_{ji}h_{ij}}{1-h_{ii}} = h_{jj} + \frac{h_{ij}^2}{1-h_{ii}} \qquad \left| h_{ij} = h_{ji} \right. .$$

(c) If $j=k=i$

$$h_{ii}^{(-i)} = h_{ii} + \frac{h_{ii}^2}{1-h_{ii}} = \frac{h_{ii} - h_{ii}^2 + h_{ii}^2}{1-h_{ii}} = \frac{h_{ii}}{1-h_{ii}} .$$

(d) We have

$$h_{jj}^{(-i)} = h_{jj} + \frac{h_{ij}{}^2}{1-h_{ii}},$$

the second term of the right side is positive.

Hence, $h_{jj}^{(-i)} \geq h_{jj}$, *i.e.*, h_{jj} is nondecreasing in n.

Property 2.4.2

Excluding the i-th observation from the analysis,

(a) $$\hat{\beta}^{(-i)} = \hat{\beta} - \frac{\left(X^T X\right)^{-1} x_i \hat{\varepsilon}_i}{1-h_{ii}}$$

(b) $$\hat{y}_i - \hat{y}^{(-i)} = x_i^T (\hat{\beta} - \hat{\beta}^{(-i)}) = \frac{h_{ii}\hat{\varepsilon}_i}{1-h_{ii}}$$

Proof:

(a) $$\hat{\beta}^{(-i)} = \left(X_{(i)}^T X_{(i)}\right)^{-1} X_{(i)}^T Y_{(i)}$$

$$= \left[\left(X^T X\right) - x_i x_i^T\right]^{-1} \left(X^T Y - x_i y_i\right)$$

$$= \left[\left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} x_i x_i^T \left(X^T X\right)^{-1}}{1 - x_i \left(X^T X\right)^{-1} x_i^T}\right] \left(X^T Y - x_i y_i\right)$$

$$= \hat{\beta} - \left(X^T X\right)^{-1} x_i y_i + \frac{\left(X^T X\right)^{-1} x_i \hat{y}_i}{1-h_{ii}} - \frac{\left(X^T X\right)^{-1} x_i h_{ii} y_i}{1-h_{ii}}$$

$$= \hat{\beta} - \frac{\left(X^T X\right)^{-1} x_i y_i (1-h_{ii}) - \left(X^T X\right)^{-1} x_i \hat{y}_i}{1-h_{ii}} - \frac{\left(X^T X\right)^{-1} x_i h_{ii} y_i}{1-h_{ii}}$$

$$= \hat{\beta} - \frac{\left(X^T X\right)^{-1} x_i y_i}{1-h_{ii}} + \frac{\left(X^T X\right)^{-1} x_i h_{ii} y_i}{1-h_{ii}} + \frac{\left(X^T X\right)^{-1} x_i \hat{y}_i}{1-h_{ii}} - \frac{\left(X^T X\right)^{-1} x_i h_{ii} y_i}{1-h_{ii}}$$

$$= \hat{\beta} - \frac{(X^T X)^{-1} x_i}{1 - h_{ii}} (y_i - \hat{y}_i) = \hat{\beta} - \frac{(X^T X)^{-1} x_i \hat{\varepsilon}_i}{1 - h_{ii}}.$$

(b) $$\hat{\beta} - \hat{\beta}^{(-i)} = \frac{(X^T X)^{-1} x_i \hat{\varepsilon}_i}{1 - h_{ii}}$$

$$\Rightarrow x_i^T (\hat{\beta} - \hat{\beta}^{(-i)}) = \frac{x_i^T (X^T X)^{-1} x_i \hat{\varepsilon}_i}{1 - h_{ii}}$$

$$\Rightarrow \hat{y}_i - \hat{y}_i^{(-i)} = \frac{h_{ii} \hat{\varepsilon}_i}{1 - h_{ii}}.$$

Property 2.4.3

After deleting i-th case from the analysis,

(a) $$\hat{\varepsilon}^{(-i)} = \hat{\varepsilon} + X(X^T X)^{-1} (1 - h_{ii})^{-1} x_i \hat{\varepsilon}_i$$

(b) $$\hat{\varepsilon}_i^{(-i)} = \hat{\varepsilon}_i \left(1 + \frac{h_{ii}}{1 - h_{ii}} \right)$$

(c) $$\hat{\varepsilon}_i^{(-i)} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}}$$

(d) $$\hat{\varepsilon}_i^{(-i)} + \hat{y}_i^{(-i)} = \hat{\varepsilon}_i + \hat{y}_i$$

Proof:

(a) $$\hat{\varepsilon}^{(-i)} = Y - \hat{Y}^{(-i)}$$

$$= Y - X\hat{\beta}^{(-i)}$$

$$= Y - X\left(\hat{\beta} - \frac{(X^T X)^{-1} x_i \hat{\varepsilon}_i}{1 - h_{ii}} \right)$$

$$= Y - X\hat{\beta} + \frac{X(X^T X)^{-1} x_i \hat{\varepsilon}_i}{1 - h_{ii}}$$

$$= \hat{\varepsilon} + X(X^T X)^{-1} (1 - h_{ii})^{-1} x_i \hat{\varepsilon}_i .$$

(b) $$\hat{\varepsilon}_i^{(-i)} = y_i - \hat{y}_i^{(-i)}$$

$$= y_i - x_i^T \hat{\beta}^{(-i)}$$

$$= y_i - x_i^T \left(\hat{\beta} - \frac{\left(X^T X\right)^{-1} x_i \hat{\varepsilon}_i}{1 - h_{ii}} \right)$$

$$= y_i - x_i^T \hat{\beta} + \frac{x_i^T \left(X^T X\right)^{-1} x_i \hat{\varepsilon}_i}{1 - h_{ii}}$$

$$= \hat{\varepsilon}_i + \frac{h_{ii} \hat{\varepsilon}_i}{1 - h_{ii}}$$

$$= \hat{\varepsilon}_i \left(1 + \frac{h_{ii}}{1 - h_{ii}} \right).$$

(c) $$\hat{\varepsilon}_i^{(-i)} = \hat{\varepsilon}_i \left(\frac{1 - h_{ii} + h_{ii}}{1 - h_{ii}} \right) = \frac{\hat{\varepsilon}_i}{1 - h_{ii}}.$$

(d) $$\hat{\varepsilon}_i^{(-i)} = \hat{\varepsilon}_i + x_i^T \left(\hat{\beta} - \hat{\beta}^{(-i)} \right)$$ [using property 2.4.3 (b) and 2.4.2 (b)]

$$= \hat{\varepsilon}_i + \hat{y}_i - \hat{y}_i^{(-i)}.$$

$$\therefore \hat{\varepsilon}_i^{(-i)} + \hat{y}_i^{(-i)} = \hat{\varepsilon}_i + \hat{y}_i.$$

Property 2.4.4

The eigen values of H and $(I\text{-}H)$ are either 0 or 1.

Proof:

Both H and $(I\text{-}H)$ are idempotent matrices. We know that the eigen values of an idempotent matrix are 0 or 1, hence the eigen values of H and $(I\text{-}H)$ are either 0 or 1.

Property 2.4.5

There are $(n\text{-}k)$ eigen values of H equal to 0, and the remaining k eigen values equal to 1.similarly, k eigen values of $(I\text{-}H)$ equal to 0 and $(n\text{-}k)$ equal to 1.

Proof:

Let λ_i, $i = 1,2,\cdots,n$ be the eigen values of H.

Since H is an idempotent matrix, we get λ_is are either 0 or 1 for all i, and

$$trace(H) = \sum_{i}^{n} \lambda_i = k .$$

Therefore k eigen values of H are must be 1 and the remaining $(n\text{-}k)$ are zero, 0.
The diagonal elements $(I\text{-}H)$ will be zero when the diagonal elements at the same position will be 1. Therefore, k eigen values of $(I\text{-}H)$ equal to 0 and $(n\text{-}k)$ eigen values equal to 1.

2.4.2 Group Deletion (Deletion of Multiple Rows)

Group deletion means deletion of multiple rows from the whole data set matrix. When a group of observations is deleted at a time we see different types of effects are on every type of analysis and estimations as follows.

Property 2.4.6 (Chattrejee and Hadi, 1988)
Suppose that X is $n \times k$ of rank k and that there are $m < k$ diagonal elements of H equal to 1. Let $I = \{i : h_{ii} = 1, \quad i = 1,2,...,n\}$ be the set of their indices. Then, for any $J \supseteq I$, $rank(X_{(J)}) \leq k - m$, with equality if J=I.

Proof:
Without loss of generality, we arrange the rows of X such that the m rows indexed by I are the last m rows of X. Thus X can be written as

$$X = \begin{pmatrix} X_{(I)} \\ X_I^T \end{pmatrix}_{m \times k}^{(n-m) \times k} .$$

Let H_I denote the principal minor of H corresponding to X_I^T. Since $h_{ii} = 1, i \in I$, then $H_I = $I, where I is the identity matrix of dimension $m \times m$. By property 2.3.6 (a) $(if\ h_{ii} = 1 \quad or\ 0, \quad then \quad h_{ij} = 0)$, H can be expressed as

$$H = \begin{pmatrix} X_{(I)}(X^T X)^{-1} X_{(I)}^T & 0 \\ 0 & I \end{pmatrix},$$

from which we see that $(X_{(I)}(X^T X)^{-1} X_{(I)}^T)$ is idempotent with rank $(k - m)$.

Hence $$(k-m)=rank\left\{X_{(I)}\left(X^TX\right)^{-1}X_{(I)}^T\right\}=rank\left\{X_{(I)}\right\}$$

It follows that, for any $J \supseteq I$,

$$rank\left(X_{(J)}\right)\le rank\left(X_{(I)}\right)=k-m.$$

Property 2.4.7

Let H_I be an $m\times m$ minor of H given by the intersection of the rows and columns of H indexed by I. If $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_m$ are the eigenvalues of H_I, then

(a) The eigenvalues of H_I and $(I-H_I)$ are between 0 and 1 inclusive; that is,

$0\le\lambda_j\le 1, j=1,2,\cdots,m.$

(b) $(I-H_I)$ is positive definite (p.d) if $\lambda_m<1$; otherwise $(I-H_I)$ is positive semidefinite (p.s.d).

Proof:

Let H_{22} be an $m\times m$ minor of H. Without loss of generality, let these be the last m rows and columns of H. Partition H as

$$H=\begin{bmatrix}H_{11} & H_{12}\\ H_{12}^T & H_{22}\end{bmatrix}.$$

Since H_{22} is symmetric, it can be written as

$$H_{22}=V\Lambda V^T, \tag{2.30}$$

where Λ is a diagonal matrix with the eigenvalues of H_{22} in the diagonal and V is a matrix containing the corresponding normalized eigenvectors as columns. Now since

$$H=HH,$$

$$H=\begin{bmatrix}H_{11} & H_{12}\\ H_{12}^T & H_{22}\end{bmatrix}=\begin{bmatrix}H_{11} & H_{12}\\ H_{12}^T & H_{22}\end{bmatrix}\begin{bmatrix}H_{11} & H_{12}\\ H_{12}^T & H_{22}\end{bmatrix}=\begin{bmatrix}- & -\\ - & H_{12}^TH_{12}+H_{22}H_{22}\end{bmatrix}$$

then $H_{22}=H_{12}^TH_{12}+H_{22}H_{22}$. Since $H_{12}^TH_{12}\ge 0$, then (a matrix $H\ge 0$ we mean $V^THV\ge 0\ \forall\ V$)

$$H_{22}\ge H_{22}H_{22}. \tag{2.31}$$

Substituting (2.30) into (2.31) gives

$$V\Lambda V^T \geq V\Lambda V^T V\Lambda V^T = V\Lambda^2 V^T,$$

or $$\Lambda \geq \Lambda^2,$$

or $$(\Lambda - \Lambda^2) \geq 0.$$

Since Λ is diagonal, hence λ_i's are diagonal elements of Λ, and

$$0 \leq \lambda_i \leq 1. \tag{2.32}$$

Part (b) If $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$ are the eigenvalues of H_{22}, then $(1-\lambda_j)$, $j = 1,2,\cdots,m$ are the eigenvalues of $(I - H_{22})$. Now, if $\lambda_m = 1$, then $(1-\lambda_m) = 0$, and hence $(I - H_{22})$ is positive semidefinite. But if $\lambda_m < 1$, then $(1-\lambda_i) > 0$, $i = 1,2,\cdots,m$, and hence $(I - H_{22})$ is positive definite.

Theorem 2.4.1 **Frisch-Waugh-Lovell (FWL)**

Consider the partitioned regression model for $K_1 + K_2$ regressors estimated in matrix form:

$$Y = [X_1 : X_2]\beta + \varepsilon$$

then

$$\hat{\beta}_2 = \left[(M_1 X_2)^T (M_1 X_2)\right]^{-1} (M_1 X_2)^T M_1 Y$$

M_1 matrix makes residuals for regression for the X_1 variables; $M_1 Y$ is the vector of residuals from regressing Y on X_1 variables. $M_1 X_2$ is the matrix made up of the column-by-column residuals of regressing each variable (column) in X_2 on all the variables in X_1.

Proof:

We have, $$(X^T X)\hat{\beta} = X^T Y$$

$$\Rightarrow \begin{bmatrix} (X_1^T X_1)\hat{\beta}_1 + (X_1^T X_2)\hat{\beta}_2 \\ (X_2^T X_1)\hat{\beta}_1 + (X_2^T X_2)\hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1^T Y \\ X_2^T Y \end{bmatrix}$$

Hence,

$$\begin{aligned}
&\left(X_1^T X_1\right)\hat{\beta}_1 + \left(X_1^T X_2\right)\hat{\beta}_2 = X_1^T Y \\
&\left(X_1^T X_1\right)\hat{\beta}_1 = X_1^T Y - \left(X_1^T X_2\right)\hat{\beta}_2 \\
&\hat{\beta}_1 = \left(X_1^T X_1\right)^{-1} X_1^T Y - \left(X_1^T X_1\right)^{-1}\left(X_1^T X_2\right)\hat{\beta}_2 \\
&\qquad = \left(X_1^T X_1\right)^{-1} X_1^T \left(Y - X_2\hat{\beta}_2\right) \qquad (2.33)
\end{aligned}$$

and

$$\left(X_2^T X_1\right)\hat{\beta}_1 + \left(X_2^T X_2\right)\hat{\beta}_2 = X_2^T Y . \qquad (2.34)$$

Putting the value of $\hat{\beta}_1$ from (2.33) in (2.34) we get

$$\begin{aligned}
&\left(X_2^T X_1\right)\left(X_1^T X_1\right)^{-1} X_1^T \left(Y - X_2\hat{\beta}_2\right) + \left(X_2^T X_2\right)\hat{\beta}_2 = X_2^T Y \\
&\Rightarrow \left(X_2^T X_1\right)\left(X_1^T X_1\right)^{-1} X_1^T Y - \left(X_2^T X_1\right)\left(X_1^T X_1\right)^{-1} X_1^T X_2\hat{\beta}_2 + \left(X_2^T X_2\right)\hat{\beta}_2 = X_2^T Y \\
&\Rightarrow \left(X_2^T X_2\right)\hat{\beta}_2 - \left(X_2^T X_1\right)\left(X_1^T X_1\right)^{-1} X_1^T X_2\hat{\beta}_2 = X_2^T Y - \left(X_2^T X_1\right)\left(X_1^T X_1\right)^{-1} X_1^T Y \\
&\Rightarrow X_2^T\left[I - X_1\left(X_1^T X_1\right)^{-1} X_1^T\right]X_2\hat{\beta}_2 = X_2^T\left[I - X_1\left(X_1^T X_1\right)^{-1} X_1^T\right]Y \\
&\Rightarrow X_2^T M_1 X_2\hat{\beta}_2 = X_2^T M_1 Y; \quad let\left(I - X_1\left(X_1^T X_1\right)^{-1} X_1^T\right) = M_1 \\
&\therefore \hat{\beta}_2 = \left(X_2^T M_1 X_2\right)^{-1} X_2^T M_1 Y \\
&\qquad = X_2^{-1} M_1^{-1}\left(X_2^T\right)^{-1} X_2^T M_1 Y \\
&\qquad = X_2^{-1} M_1^{-1} M_1 Y = \left(M_1 X_2\right)^{-1} M_1 Y \\
&\qquad = \left(M_1 X_2\right)^{-1}\left[\left(M_1 X_2\right)^T\right]^{-1}\left(M_1 X_2\right)^T M_1 Y \\
&\qquad = \left[\left(M_1 X_2\right)^T\left(M_1 X_2\right)\right]^{-1}\left(M_1 X_2\right)^T M_1 Y .
\end{aligned}$$

M_1 is both idempotent and symmetric , we can then rewrite as,

$$\hat{\beta}_2 = \left(X_2^{*T} X_2^{*}\right)^{-1} X_2^{*T} y^{*} . \qquad \text{where, } \begin{matrix} X_2^{*} = M_1 X_2 \\ y^{*} = M_1 Y \end{matrix}$$

Property 2.4.8

If X is partitioned as $X = (X_1 : X_2)$ *and* X_1 *and* X_2 are two sets of independent explanatory variables. Then the prediction matrix H can be represented as,

$$H = H_1 + H_2$$
$$= X_1(X_1^T X_1)^{-1} X_1^T + X_2(X_2^T X_2)^{-1} X_2$$

Proof:

$$H = X\left(X^T X\right)^{-1} X^T$$
$$= \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix}$$

We have two sets of independent variables, that are orthogonal to each other, and then the sums of cross products of the variables in X_1 with X_2 are zero by definition. Thus $(X^T X)$ matrix formed out of X_1 and X_2 is block diagonal. Hence

$$H = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{bmatrix} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix}$$
$$= \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} (X_1^T X_1)^{-1} & 0 \\ 0 & (X_2^T X_2)^{-1} \end{bmatrix} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix}$$
$$= \begin{bmatrix} X_1\left(X_1^T X_1\right)^{-1} & X_2\left(X_2^T X_2\right)^{-1} \end{bmatrix} \begin{bmatrix} X_1^T \\ X_2^T \end{bmatrix}$$
$$= X_1\left(X_1^T X_1\right)^{-1} X_1^T + X_2\left(X_2^T X_2\right)^{-1} X_2^T$$
$$= H_1 + H_2 .$$

2.5 Group Deletion Algebra

Property 2.5.1

If more than one observation (a group contains d observations) are deleted then the idea of group deletion (GD) has been introduced.

i) $\hat{\beta}^{(-D)} = \hat{\beta} - \left(X^T X\right)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D$

ii) The GD residual, $\hat{\varepsilon}^{(-D)}$ is defined as,

$$\hat{\varepsilon}^{(-D)} = Y - X\hat{\beta}^{(-D)}$$

$$\hat{\varepsilon}^{(-D)} = \hat{\varepsilon} + X\left(X^T X\right)^{-1} X_D^T \left(I_D - U_D\right)^{-1} \hat{\varepsilon}_D$$

iii) The sum of squared differences of $\hat{\varepsilon}^{(-D)}$ for $\hat{\varepsilon}^{(-D)}$ is

$$\left(\hat{\varepsilon}^{(-D)} - \hat{\varepsilon}\right)^T \left(\hat{\varepsilon}^{(-D)} - \hat{\varepsilon}\right) = \hat{\varepsilon}_D{}^T \left(I_D - U_D\right)^{-1} U_D \left(I_D - U_D\right)^{-1} \hat{\varepsilon}_D$$

iv) $\hat{\varepsilon}^{(-D)}$ an be partitioned as,

$$\hat{\varepsilon}^{(-D)} = \begin{bmatrix} \hat{\varepsilon}_R{}^{(-D)} \\ \hat{\varepsilon}_D{}^{(-D)} \end{bmatrix} = \begin{bmatrix} \hat{\varepsilon}_R + V\left(I_D - U_D\right)^{-1} \hat{\varepsilon}_D \\ \left(I_D - U_D\right)^{-1} \hat{\varepsilon}_D \end{bmatrix}$$

Proof:

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-d*) cases after *d*< *(n-p)* cases in *D* are deleted. Without loss of generality, assume that these observations are the last *d* rows of *X*, *Y* and ε so that they can be partitioned as

$$X = \begin{bmatrix} X_R & (n-d)\times p \\ X_D & d\times p \end{bmatrix}, \; Y = \begin{bmatrix} Y_R & (n-d)\times 1 \\ Y_D & d\times 1 \end{bmatrix}, \; \varepsilon = \begin{bmatrix} \varepsilon_R & (n-d)\times 1 \\ \varepsilon_D & d\times 1 \end{bmatrix}$$

$$H = \begin{bmatrix} X_R\left(X^T X\right)^{-1} X_R^T & X_R\left(X^T X\right)^{-1} X_D^T \\ X_D\left(X^T X\right)^{-1} X_R^T & X_D\left(X^T X\right)^{-1} X_D^T \end{bmatrix} = \begin{bmatrix} U_R & V \\ V^T & U_D \end{bmatrix} .$$

The weight or prediction matrix $H = X\left(X^T X\right)^{-1} X^T$ can be partitioned as above,

where $U_R = X_R\left(X^T X\right)^{-1} X_R^T$ and $U_D = X_D\left(X^T X\right)^{-1} X_D^T$ are $(n-d)\times(n-d)$ *and* $d\times d$ symmetric matrices, and $V = X_R\left(X^T X\right)^{-1} X_D^T$ is an $(n-d)\times d$, matrix.

(i) $\hat{\beta}^{(-D)} = \left(X_R^T X_R\right)^{-1} X_R^T Y_R$

$$= \left(X^T X - X_D^T X_D\right)^{-1} X_R^T Y_R$$

$$= \left[\left(X^T X\right)^{-1} + \left(X^T X\right)^{-1} X_D^T \left(I_D - X_D\left(X^T X\right)^{-1} X_D^T\right)^{-1} X_D\left(X^T X\right)^{-1}\right]\left(X^T Y - X_D^T Y_D\right)$$

$$= (X^T X)^{-1} X^T Y + (X^T X)^{-1} X_D^T \left(I_D - X_D (X^T X)^{-1} X_D^T\right)^{-1} X_D (X^T X)^{-1} X^T Y$$
$$- (X^T X)^{-1} X_D^T Y_D - (X^T X)^{-1} X_D^T \left(I_D - X_D (X^T X)^{-1} X_D^T\right)^{-1} X_D (X^T X)^{-1} X_D^T Y_D$$
$$= \hat{\beta} + (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} X_D \hat{\beta} - (X^T X)^{-1} X_D^T Y_D - (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} U_D Y_D$$
$$= \hat{\beta} + (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} X_D \hat{\beta} - (X^T X)^{-1} X_D^T \left[I_D + (I_D - U_D)^{-1} U_D\right] Y_D$$
$$= \hat{\beta} + (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} X_D \hat{\beta} - (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} Y_D$$
$$= \hat{\beta} - (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \left[Y_D - X_D \hat{\beta}\right]$$
$$= \hat{\beta} - (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D .$$

(ii) The GD residual vector, $\hat{\varepsilon}^{(-D)}$ can be expressed in terms of the LS residual vector $\hat{\varepsilon}$.

$$\hat{\varepsilon}^{(-D)} = Y - X\hat{\beta}^{(-D)}$$
$$= Y - X\left(\hat{\beta} - (X^T X)^{-1} X_D^T (I_D - U_p)^{-1} \hat{\varepsilon}_D\right)$$
$$= Y - X\hat{\beta} + X(X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D$$
$$= \hat{\varepsilon} + X(X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D .$$

(iii) $$\left(\hat{\varepsilon}^{(-D)} - \hat{\varepsilon}\right) = X(X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D ,$$

hence the sum of squared differences of $\hat{\varepsilon}^{(-D)}$ *from* $\hat{\varepsilon}$ *is*

$$\therefore \left(\hat{\varepsilon}^{(-D)} - \hat{\varepsilon}\right)^T \left(\hat{\varepsilon}^{(-D)} - \hat{\varepsilon}\right) = \hat{\varepsilon}_D^T (I_D - U_D)^{-1} X_D (X^T X)^{-1} X^T X (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D$$
$$= \hat{\varepsilon}_D^T (I_D - U_D)^{-1} X_D (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D$$
$$= \hat{\varepsilon}_D^T (I_D - U_D)^{-1} U_D (I_D - U_D)^{-1} \hat{\varepsilon}_D .$$

(iv) $$\hat{\varepsilon}^{(-D)} = \begin{bmatrix} \hat{\varepsilon}_R^{(-D)} \\ \hat{\varepsilon}_D^{(-D)} \end{bmatrix} = \begin{bmatrix} Y_R - \hat{Y}_R^{(-D)} \\ Y_D - \hat{Y}_D^{(-D)} \end{bmatrix} = \begin{bmatrix} Y_R - X_R \hat{\beta}^{(-D)} \\ Y_D - X_D \hat{\beta}^{(-D)} \end{bmatrix}$$

and

$$\hat{\varepsilon}_R^{(-D)} = Y_R - X_R \hat{\beta}^{(-D)}$$
$$= Y_R - X_R \left(\hat{\beta} - (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D\right)$$
$$= Y_R - X_R \hat{\beta} + X_R (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D$$
$$= \hat{\varepsilon}_R + V(I_D - U_D)^{-1} \hat{\varepsilon}_D$$

$$\hat{\varepsilon}_D^{(-D)} = Y_D - X_D \hat{\beta}^{(-D)}$$
$$= Y_D - X_D \left[\hat{\beta} - (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D\right]$$
$$= Y_D - X_D \hat{\beta} + X_D (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\varepsilon}_D$$

$$= \hat{\varepsilon}_D + U_D (I_D - U_D)^{-1} \hat{\varepsilon}_D$$
$$= \left[I_D + U_D (I_D - U_D)^{-1} \right] \hat{\varepsilon}_D$$
$$= (I_D - U_D)^{-1} \hat{\varepsilon}_D .$$

Property 2.5.2 (Imon, 1996)

(i) $Rank(U_R) = p$ (ii) $Rank(U_D) = \min(d, p)$

(iii) $Trace(U_R) = \sum_{i \in R} h_{ii}$ (iv) $Trace(U_D) = \sum_{i \in D} h_{ii}$

Proof:

We observe from the consideration of U_R and U_D matrices that they are symmetric. But they are not idempotent, though they are sub matrices of the idempotent matrix H. As we know, X is an $n \times p$ matrix of rank p, X_R is an $(n-d) \times p$ matrix of rank p, and $X^T X$ is a $p \times p$ nonsingular matrix.

$U_R = X_R (X^T X)^{-1} X_R^T$ is an $(n-d) \times (n-d)$ matrix of rank p;

i.e., $Rank(U_R) = p$, assuming $p < n - d$.

On the other hand, X_D is a $d \times p$ matrix of rank = *min* (d, p), and hence $U_D = X_D (X^T X)^{-1} X_D^T$ is a $d \times d$ matrix having rank = *min* (d, p);

i.e., $Rank(U_D) = \min(d, p)$.

From the partitioned form of H we get,

$$Trace(H) = Trace[U_R + U_D]$$
$$= Trace(U_R) + Trace(U_D),$$

where $Trace(U_R) = \sum_{i \in R} h_{ii}$ *and*, $Trace(U_D) = \sum_{i \in D} h_{ii}$.

Idempotency of the Prediction Matrix

Here we see the idempotency on the partitioned matrices.

Property 2.5.3 (Imon, 1996)

(i) $U_R U_R + VV^T = U_R$ (ii) $U_D U_D + VV^T = U_D$

(iii) $U_R V + V U_D = V$ (iv) $V^T U_R + U_D V^T = V^T$

Proof:

Since H is idempotent, $H = H^2 = H \cdot H$

$$\therefore H = \begin{pmatrix} U_R & V \\ V^T & U_D \end{pmatrix} = \begin{pmatrix} U_R & V \\ V^T & U_D \end{pmatrix}\begin{pmatrix} U_R & V \\ V^T & U_D \end{pmatrix}$$

$$= \begin{pmatrix} U_R U_R + VV^T & U_R V + VU_D \\ V^T U_R + U_D V^T & V^T V + U_D U_D \end{pmatrix}.$$

That comes to the results as,

$$U_R U_R + VV^T = U_R, \quad U_D U_D + V^T V = U_D$$

$$U_R V + VU_D = V, \qquad V^T U_R + U_D V^T = V^T.$$

Property 2.5.4 (Imon, 1996)

If **1** is a unit vector of order n and 0 is a same order null vector, then

(i) $U_R 1_R + V 1_D = 1_R$ (ii) $V^T 1_R + U_D 1_D = 1_D$

(iii) $U_R \hat{\varepsilon}_R + V\hat{\varepsilon}_D = 0_R$ (iv) $V^T \hat{\varepsilon}_R + U_D \hat{\varepsilon}_D = 0_D$

Proof:

By partitioning the relationships $H1 = 1$ *and* $H\hat{\varepsilon} = 0$,
we obtain

$$H1 = 1$$

$$= \begin{pmatrix} U_R & V \\ V^T & U_D \end{pmatrix}\begin{pmatrix} 1_R \\ 1_D \end{pmatrix} = \begin{pmatrix} U_R 1_R + V 1_D \\ V^T 1_R + U_D 1_D \end{pmatrix} = \begin{pmatrix} 1_R \\ 1_D \end{pmatrix}$$

hence
$$\begin{aligned} U_R 1_R + V 1_D &= 1_R \\ V^T 1_R + U_D 1_D &= 1_D \end{aligned}.$$

Now

$$H\hat{\varepsilon} = 0$$

i.e.,
$$= \begin{pmatrix} U_R & V \\ V^T & U_D \end{pmatrix}\begin{pmatrix} \hat{\varepsilon}_R \\ \hat{\varepsilon}_D \end{pmatrix} = \begin{pmatrix} U_R \hat{\varepsilon}_R + V\hat{\varepsilon}_D \\ V^T \hat{\varepsilon}_R + U_D \hat{\varepsilon}_D \end{pmatrix} = \begin{pmatrix} 0_R \\ 0_D \end{pmatrix}$$

hence
$$\begin{aligned} U_R \hat{\varepsilon}_R + V\hat{\varepsilon}_D &= 0_R \\ V^T \hat{\varepsilon}_R + U_D \hat{\varepsilon}_D &= 0_D \end{aligned}.$$

Variance-Covariance Pattern of Group-Deleted (GD) Residuals

Property 2.5.5 (Imon, 1996)

It is possible to express the different components of GD residual vector $\hat{\varepsilon}^{(-D)}$ in terms of the vectors, ε_R and ε_D.

Proof: As we know,

$$\begin{aligned}\hat{\varepsilon}_D &= Y_D - X_D\hat{\beta} \\ &= \varepsilon_D - X_D\left(X^TX\right)^{-1}\left(X_R^T\varepsilon_R + X_D^T\varepsilon_D\right) \\ &= \left(I_D - U_D\right)\varepsilon_D - V^T\varepsilon_R\end{aligned} \tag{2.35}$$

and substituting this result in property 2.5.1 (iv) ($\hat{\varepsilon}_D^{(-D)} = \left(I_D - U_D\right)\hat{\varepsilon}_D$) we obtain

$$\hat{\varepsilon}_D^{(-D)} = \varepsilon_D - \left(I_D - U_D\right)^{-1}V^T\varepsilon_R, \tag{2.36}$$

which also implies

$$E\left(\hat{\varepsilon}_D^{(-D)}\right) = 0_D. \tag{2.37}$$

Using the result of Henderson and Searle (1981) we also observed that

$$I_D + \left(I_D - U_D\right)^{-1}U_D = \left(I_D - U_D\right)^{-1} \tag{2.38}$$

Hence using property 2.5.3 ($U_DU_D + V^TV = U_D$) and equation 2.38

$I_D + (I_D - U_D)^{-1}U_D = (I_D - U_D)^{-1}$ we obtain

$$\begin{aligned}Var\left(\hat{\varepsilon}_D^{(-D)}\right) &= E\left(\hat{\varepsilon}_D^{(-D)}\hat{\varepsilon}_D^{(-D)T}\right) \\ &= \sigma^2\left[I_D + \left(I_D - U_D\right)^{-1}V^TV\left(I_D - U_D\right)^{-1}\right] \\ &= \sigma^2\left[I_D + \left(I_D - U_D\right)^{-1}U_D\right] \\ &= \sigma^2\left(I_D - U_D\right)^{-1}\end{aligned} \tag{2.39}$$

again

$$\begin{aligned}\hat{\varepsilon}_R &= Y_R - X_R\hat{\beta} \\ &= \left(I_R - U_R\right)\varepsilon_R - V\varepsilon_D\end{aligned} \tag{2.40}$$

using the results (2.35) and (2.40) in property 2.5.1 (iv), we obtain

$$\hat{\varepsilon}_R^{(-D)} = \left[I_R - U_R - V\left(I_D - U_D\right)^{-1}V^T\right]\varepsilon_R \tag{2.41}$$

which also implies $E\left(\hat{\varepsilon}_R^{(-D)}\right) = 0_R$. (2.42)

It is obvious from standard least squares theory that $\left[I_R - U_R - V\left(I_D - U_D\right)^{-1}V^T\right]$ must be a symmetric idempotent matrix, which implies

$$\begin{aligned} Var(\hat{\varepsilon}_R^{(-D)}) &= E(\hat{\varepsilon}_R^{(-D)}\hat{\varepsilon}_R^{(-D)T}) \\ &= \sigma^2\left[I_R - U_R - V\left(I_D - U_D\right)^{-1}V^T\right]. \end{aligned} \quad (2.43)$$

Using property 2.5.3 of U_R and U_D matrices we also observe

$$\begin{aligned} Cov\left(\hat{\varepsilon}_R^{(-D)}, \hat{\varepsilon}_D^{(-D)}\right) &= E\left(\hat{\varepsilon}_R^{(-D)}\hat{\varepsilon}_D^{(-D)T}\right) \\ &= \sigma^2\left[V\left(I_D - U_D\right)^{-1} - V\left(I_D - U_D\right)^{-1}\right] \\ &= 0_R. \end{aligned}$$

Property 2.5.6

$$H^{(-D)} = \begin{bmatrix} H_R^{(-D)} \\ H_D^{(-D)} \end{bmatrix} = \begin{bmatrix} U_R + V\left(I_D - U_D\right)^{-1}V^T \\ U_D\left(I_D + \left(I_D - U_D\right)^{-1}U_D\right) \end{bmatrix}$$

Proof:

$$\begin{aligned} H_R^{(-D)} &= X_R\left(X_R^T X_R\right)^{-1}X_R^T \\ &= X_R\left(X^T X - X_D^T X_D\right)^{-1}X_R^T \\ &= X_R\left[\left(X^T X\right)^{-1} + \left(X^T X\right)^{-1}X_D^T\left(I_D - X_D\left(X^T X\right)^{-1}X_D^T\right)^{-1}X_D\left(X^T X\right)^{-1}\right]X_R^T \\ &= X_R\left(X^T X\right)^{-1}X_R^T + X_R\left(X^T X\right)^{-1}X_D^T\left(I_D - U_D\right)^{-1}X_D\left(X^T X\right)^{-1}X_R^T \\ &= U_R + V\left(I_D - U_D\right)^{-1}V^T, \end{aligned}$$

and

$$\begin{aligned} H_D^{(-D)} &= X_D\left(X_R^T X_R\right)^{-1}X_D^T \\ &= X_D\left(X^T X - X_D^T X_D\right)^{-1}X_D^T \\ &= X_D\left[\left(X^T X\right)^{-1} + \left(X^T X\right)^{-1}X_D^T\left(I_D - U_D\right)^{-1}X_D\left(X^T X\right)^{-1}\right]X_D^T \\ &= U_D + U_D\left(I_D - U_D\right)^{-1}U_D \\ &= U_D\left(I_D + \left(I_D - U_D\right)^{-1}U_D\right). \end{aligned}$$

Property 2.5.7

$$h_{ii}^{(-D)} \geq h_{ii}$$

Proof:

We have, $$h_{ii}^{(-D)} = x_i^T \left(X_R^T X_R\right)^{-1} x_i$$

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

$$= x_i^T \left[\left(X^T X\right)^{-1} + \left(X^T X\right)^{-1} X_D^T (I_D - U_D)^{-1} X_D \left(X^T X\right)^{-1}\right] x_i$$

$$= x_i^T \left(X^T X\right)^{-1} x_i + x_i^T \left(X^T X\right)^{-1} X_D^T (I_D - U_D)^{-1} X_D \left(X^T X\right)^{-1} x_i .$$

$$= h_{ii} + x_i^T \left(X^T X\right)^{-1} X_D^T (I_D - U_D)^{-1} X_D \left(X^T X\right)^{-1} x_i$$

Hence $$h_{ii}^{(-D)} \geq h_{ii} .$$

2.6 Deletion Diagnostics Algebra in Existing Methods

This section provides some basic relationships among the deletion diagnostic measures.

Property 2.6.1

(i) $$CD_i = \frac{e_i^2 h_{ii}}{p(1-h_{ii})} \tag{2.44}$$

(ii) $$DFFITS_i = r_i^* \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \tag{2.45}$$

Proof:

(i) We have

$$CD_i = \frac{1}{ps^2}\left[\left\{X\left(\hat{\beta} - \hat{\beta}^{(-i)}\right)\right\}^T \left\{X\left(\hat{\beta} - \hat{\beta}^{(-i)}\right)\right\}\right]$$

$$= \frac{1}{ps^2}\left[\left(\hat{\beta} - \hat{\beta}^{(-i)}\right)^T X^T X\left(\hat{\beta} - \hat{\beta}^{(-i)}\right)\right]$$

$$= \frac{1}{ps^2}\left[\frac{\hat{\varepsilon}_i x_i^T \left(X^T X\right)^{-1}}{1-h_{ii}} \left(X^T X\right) \frac{\left(X^T X\right)^{-1} x_i \hat{\varepsilon}_i}{1-h_{ii}}\right] ,$$

putting the value of $\left(\hat{\beta}-\hat{\beta}^{(-i)}\right)$, property 2.4.2

$$=\frac{h_{ii}\hat{\varepsilon}_i^2}{ps^2(1-h_{ii})^2},$$

since $\hat{\varepsilon}_i x_i^T\left(X^TX\right)^{-1}x_i\hat{\varepsilon}_i=\hat{\varepsilon}_i^2 h_{ii}$

$$=\frac{h_{ii}}{p(1-h_{ii})}\frac{\hat{\varepsilon}_i^2}{s^2(1-h_{ii})} \qquad \left| e_i^2=\frac{\hat{\varepsilon}_i^2}{s^2(1-h_{ii})}\right.$$

$$=\frac{e_i^2 h_{ii}}{p(1-h_{ii})}.$$

(ii) We Know

$$DFFITS_i=\frac{y_i-\hat{y}^{(-i)}}{\sqrt{s_{(i)}^2 h_{ii}}},\ i=1,2,\ldots,n$$

and we have $\hat{\beta}_i-\hat{\beta}^{(-i}=\frac{\left(X^TX\right)^{-1}x_i\hat{\varepsilon}_i}{1-h_{ii}}.$

Multiplying both sides by x_i^T produces

$$y_i-\hat{y}^{(-i)}=\frac{h_{ii}\hat{\varepsilon}_i}{1-h_{ii}}.$$

Dividing both sides by $\sqrt{s_{(i)}^2 h_{ii}}$ will produce $DFFITS_i$:

$$DFFITS_i=\frac{y_i-\hat{y}^{(-i}}{\sqrt{s_{(i)}^2 h_{ii}}}=\frac{h_{ii}\hat{\varepsilon}_i}{1-h_{ii}}\left[\frac{1}{s_{(i)}^2 h_{ii}}\right]^{1/2}$$

$$=\frac{\hat{\varepsilon}_i}{\sqrt{s_{(i)}^2(1-h_{ii})}}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

$$=r_i^{\bullet}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2},$$

where $r_i^{\bullet}$ is the R-Student residual.

2.7 Sources and Nature of Data

We use two types of data in our dissertation; one is secondary data and the other is simulated data. Secondary data are collected from different referred books and journals that are extensively used in diagnostic purpose and number of statisticians use them to evaluate the performance and effectiveness of their own measures. We also use simulated data for the purpose of simulation of the proposed measures. We generate and use simulated data to demonstrate the performance of our measures for the case of high-dimensional large data set. All the secondary data sets are given in appendix A.

Chapter 3

"Finding the question is often more important than finding the answer."

John W. Tukey

Robust Regression and Regression Diagnostics: Deletion Approach

This chapter provides a short literature review with the art of the state of robust regression and regression diagnostics. This chapter is divided into two sections; in the first section we make a short discussion about most popular robust regression methods and a very short introduction to the robustness properties. Next section contains different regression diagnostic methods that are mainly based on the deletion approach.

3.1 Robust Regression

This section gives the concepts of robust regression, measures of robustness and some popular robust regression methods.

3.1.1 Main Concepts in Robust Regression

Robust regression techniques that are complement to the classical least squares (LS) in the sense that they give similar results to the LS regression when the data are linear with normally distributed errors, but differ significantly when the errors are non normal or the data set contains significant outliers. Robust regression tries to fit a regression to majority of the data and then to discover the outliers as those points which posses' large residuals from the robust output. These have the common strategy; all give less weight to

observations that would otherwise influence the regression line. Robust regressions are considered as good or better depend upon their robustness properties.

3.1.2 Measures of Robustness

Three most common measures of assessing the robustness of an estimator T (F_n) are the breakdown point, influence function and continuity of the descriptive measure T which induces it. These notions are often described as quantitative, infinitesimal and qualitative robustness. Very short discussions of the three robustness properties: breakdown point, influence function and asymptotic normality are as follows.

The breakdown point characterizes the maximal deviation (in the sense of metric chosen) from the ideal model F_0 that provides the boundness of the estimator bias.

Breakdown point, as applied to the Huber supermodel or gross-error model

$$\varepsilon^*(T,F)=\sup_{\varepsilon<1}\left\{\varepsilon:\sup_{F:F=(1-\varepsilon)F_0+\varepsilon H}\left|T(F)-T(F_0)\right|<\infty\right\}. \quad (3.1.1)$$

This notion defines the largest fraction of gross errors that still keeps the bias bounded. In regression T is *(p+2)* dimensional functional defined on *(p+1)* dimensional sample space. Two most important sample based classification of breakdown point are, additional breakdown point (Hampel *et al.* 1986) and replacement breakdown point

Let, $x^n=\{x_1,x_2,\ldots,x_n\}$ of size n in R^p = the finite sample addition breakdown point of an estimator is defined as

$$ABP(T,X^n)=\min\left\{\frac{m}{m+n}:\sup\left\|T(x^n\cup y^n)-T(x^n)\right\|=\infty\right\}, \quad (3.1.2)$$

where y^n denotes a data set of size m with arbitrary values, and $x^n\cup y^m$ denotes contaminated sample by adjoining y^m to x^n .The finite sample replacement breakdown point of an estimates T at x^n is defined as

$$RBP(T,X^n)=\min\left\{\frac{m}{n}:\sup_{x_m^n}\left\|T(x_m^n)-T(x^n)\right\|=\infty\right\} \quad (3.1.3)$$

where x_m^n denotes one corrupted sample from x^n by replacing m points of with x^n arbitrary values.

Influence Function (IF) assess the robustness concept of an estimator T (F_n), it (IF) is the cornerstone of the infinitesimal approach, which was invented by Hampel (1968). It gives us a precise idea of how the estimator responds to a small amount of contamination (infinitesimal perturbation) at any point. The influence function at z can be thought of as an approximation to the relative change in an estimate caused by the addition of a small proportion of spurious observations at z.

The influence function (IF) of T at F is given by

$$IF(z;T,F) = \lim_{t\downarrow 0} \frac{T((1-t)F + t\Delta_z) - T(F)}{t} \tag{3.1.4}$$

in those $Z \in \chi$ where this limit exists. In regression, $Z = (Y : X)$; X is a p-dimensional vector of regressors and Y is a one-dimensional vector of response variable. Hampel *et al.* (1982) defines two versions of finite sample influence functions, one by addition and the other by replacement of observations.

Asymptotic Normality (univariate), A sequence of random variables $\{X_n\}$ converges in distribution to $N(\mu,\sigma^2)$, $\sigma > 0$, if equivalently, the sequence $\{(x_n - \mu)/\sigma\}$ converges in distribution to $N(0,1)$. More generally, a sequence of random variables $\{X_n\}$ is asymptotically normal with "mean" μ_n and "variance "σ_n^2 if $\sigma_n > 0$ for all n sufficiently large and

$$\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1). \tag{3.1.5}$$

We write "X_n"is an $AN(\mu_n,\sigma_n^2)$, where $\{\mu_n\}$ and $\{\sigma_n\}$ are sequence of constants.

Asymptotic Normality (multivariate), a sequence of random vectors $\{X_n\}$ is asymptotically (multivariate) normal with "mean vector" μ_n and "covariance matrix" $\sum_n$ if $\sum_n$ has nonzero diagonal elements for all n sufficiently large, and for every vector λ such that $\lambda \sum_n \lambda^T > 0$ for all n sufficiently large, the sequence λX_n^T

is $AN(\lambda\mu_n^T, \lambda\sum_n \lambda^T)$. We write "$X_n$ is AN (μ_n, $\sum_n$)." Here $\{\mu_n\}$ is a sequence of vector constants and $\{\sum_n\}$ a sequence of covariance matrix constants.

3.1.3 Different Robust Regression

Many types of robust regression methods have been developed in literature, some of the most popular robust regression methods are:

L_1 Regression

A first step toward a more robust regression estimator came from Edgeworth (1887) after the little improvement of the proposal of Boscovich. He argued that outliers have a very large influence on LS because the residuals r_i are squared. Therefore, he proposed the least absolute value regression estimator, which is defined as

$$\min_{\beta} imize \sum_{i=1}^{n} |r_i| \tag{3.1.6}$$

L_1 regression does not protect against outlying x, but protects against y and is quite preferable over LS in this respect. The breakdown point of L_1 is 0%.

Robust M-Estimator

Huber (1973) introduced M estimator in regression, that he had developed in 1964 to estimate location parameter robustly. The name "M-estimator" (Huber, 1964) comes from "generalized maximum likelihood". Robust M-estimators attempt to limit the influence of outliers and based on the idea of replacing the squared residuals r_i^2 used in LS estimation with less rapidly increasing loss-function of the data value and parameter estimate, yielding

$$\min imize \sum_{i=1}^{n} \rho(r_i), \tag{3.1.7}$$

where ρ is a symmetric, positive-definite function generally with a unique minimum at zero. Differentiating this expression with respect to the regression coefficients β yields

$$\sum_{i=1}^{n} \psi(r_i)\mathbf{x}_i = 0 \tag{3.1.8}$$

where ψ, the derivative of ρ, and $\mathbf{x}_i$ is the row vector of explanatory variables of the i-th case :

$$\mathbf{x}_i = (x_{i1},...,x_{ip})$$
$$\mathbf{0} = (0,...,0).$$

Equation (3.1.8) is a system of linear equation. One uses iteration schemes based on reweighted LS (Holland and Welsch 1977) or the so-called H-algorithm (Huber and Dutter 1974, Dutter 1977, Marazzi 1980). Unlike equation 3.1.6, however the solution of 3.1.8 is not equivariant with respect to a magnification of the y-axis. Therefore, one has to standardize the residuals by means of some estimate of σ, yielding

$$\sum_{i=1}^{n} \psi\left(\frac{r_i}{\hat{\sigma}}\right) x_i = \mathbf{0}, \tag{3.1.9}$$

where these must be estimated simultaneously. Motivated by minimax asymptotic variance arguments, Huber proposed to use the function

$$\psi(t) = \min(c, \max(t, -c)). \tag{3.1.10}$$

M-estimators are statistically more efficient (at a model with Gaussian errors) than L_1-regression while at the same time they are still robust with respect to outlying y_i. It has finite breakdown point $1/n$ due to outlying x_i.

Generalized M-estimators

In order to avoid the vulnerability to leverage points generalized M-estimators (GM-estimators) were introduced, with the basic purpose of bounding the influence of outlying x_i by means of some weight function w. Mallows (1975) suggested to replace (3.1.8) by

$$\sum_{i=1}^{n} w(\mathbf{x}_i)\psi(r_i / \hat{\sigma})\mathbf{x}_i = \mathbf{0}, \tag{3.1.11}$$

GM estimator has breakdown point *1/(p+1)*. These estimators were constructed in the hope of bounding the influence of a single outlying observation. The effect of which can be measured by means of the so-called influence function (Hampel, 1974). Therefore, the corresponding GM-estimators are generally called bounded-influence estimators.

Redescending M-estimators

If the derivative of score function ρ of an M-estimator's is redescending ψ, *i.e.* satisfies,

$$\lim_{z\Rightarrow\pm\alpha} \rho'(z) = 0, \tag{3.1.12}$$

then the M-estimator is called a redescending M-estimator. Redescending M-estimators for regression parameters have special robustness properties. Due to Donoho and Huber (1983) breakdown point of regression estimators allow outliers in the observations as well as the regressors. Maronna, Bustos and Yohai (1979) found under this definition, all M-estimators with non-decreasing ψ as the L_1 estimator behaves as bad as the least squares estimator. Later, He *et al.* (1990) and Ellis and Morgenthaler (1992) found that the situation changes completely if outliers appear only in the observations and not in the regressors, a situation which in particular appears in designed experiments where the regressors given by the experimenter.

R-estimators of Regression

R-estimators are based on the rank of the residuals, r_i. If R_i is the rank of $r_i = y_i - \hat{y}_i$, then the objective is to

$$\min imize \sum_{i=1}^{n} a_n(R_i) r_i, \tag{3.1.13}$$

where the score's function $a_n(i)$ is monotone and satisfies $\sum_{i=1}^{n} a_n(i) = 0$.

S-estimators

Rousseeuw and Yohai (1984) introduced S-estimators; the idea is to define T_n as the set of parameters β that produces residuals with the smallest dispersion. *i.e.*,

$$T_n = \arg \quad \min imize_{\beta} s(r_1(\beta),...,r_n(\beta)). \tag{3.1.14}$$

It is a certain type of robust M-estimates, $S_n(\beta)$, of the residuals $(r_1(\hat{\beta}),...,r_n(\hat{\beta}))$ is defined as the solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{s_n(\beta)}\right) = k, \tag{3.1.15}$$

k is often put equal to $E_{\phi}(\rho)$, where φ is the standard normal. s_n given in (3.1.15) is an M-estimator of scale. The function ρ must satisfy the following conditions:

(*S1*) ρ is symmetric and continuously differentiable, and $\rho(0)=0$.

(*S2*) There exists $c > 0$ such that ρ is strictly increasing on [0, c] and constant on [c,∞).

[If there happens to be more than one solution to (3.1.15), then put $s(r_1,...,r_n)$ equal to the supremum of the set of solutions; this means $s(r_1,...,r_n) = \sup\{s;(1/n)\sum\rho(r_i/s)=k\}$. If there exists no solution to (3.1.15), then put $s(r_1,...,r_n)=0$].

S-estimates for regression are consistent for the true regression parameter β and asymptotically normal when the distribution of the errors is symmetric around zero. But these estimates can not achieve high efficiency and high breakdown point at the same time.

MM-Estimators

Yohai (1985) introduced a new class of estimators named as MM-estimators toward higher efficiency for high-breakdown estimators. It is defined in three stages. In the first stage, a high-breakdown estimate β^* is calculated, such as LMS or LTS (*see* below). For this purpose, the robust estimator does not need to be efficient. Then, an M-estimate of scale s_n with 50% breakdown is computed on the residuals $r_i(\beta^*)$ from the robust fit. Finally, the MM estimator $\hat{\beta}$ is defined as any solution of

$$\sum_{i=1}^{n}\psi(r_i(\beta)/s_n)\mathbf{x}_i = \mathbf{0}, \tag{3.1.16}$$

which satisfies

$$S(\beta) \le S(\beta^*), \tag{3.1.17}$$

where

$$S(\beta) = \sum_{i=1}^{n}\rho(r_i(\beta)/s_n). \tag{3.1.18}$$

The function ρ must be like those used in the construction of S-estimators, it must satisfy conditions *S1* and *S2* of S-estimators. This implies that $\psi = \rho'$ has to be properly redescending. Yohai (1985) showed that MM-estimators inherit 50% breakdown point of the first stage and that they also possess the exact fit property, he also proved that MM-estimators are highly efficient when the errors are normally distributed.

Least Median of Squares (LMS) Regression

It was proposed by Hampel (1975) and further developed by Rousseeuw (1984). Instead of minimizing the sum of squared residuals, Rousseeuw proposed minimizing their median

$$\underset{\beta}{\text{minimize}} \; \underset{i}{med}\, r_i^2 \; . \tag{3.1.19}$$

This estimator effectively trims almost the half (*n*/2) observations having the largest residuals, and uses the maximal residual value in the remaining set as the criterion to be minimized. The basic re-sampling algorithm for approximating the LMS, called PROGRESS, was proposed by Rousseeuw and Leroy in 1987 and later developed by Rousseeuw and Hubert in 1997. This algorithm considers a trial subset of p (number of explanatory variables) observations and calculated the linear fit passing through them. This procedure is repeated many times, and the fit with the lowest median of squared residuals is retained, for small data it is possible to consider all p-subsets whereas for large data sets many p-subsets are drawn at random. The currently fastest exact algorithm by Agullo (1996) is based on a branch and bound procedure that selects the optimal h-subset without requiring the inspection of all h-subsets. This algorithm is feasible for n up to about 100 and p up to about 5. LMS has breakdown point of $([n/2]-p+2)/n)$ for p-dimensional data set *i.e.*, it attains maximum possible breakdown point *1/2* at usual models but unfortunately it possesses poor asymptotic efficiency. LMS has excellent global robustness. LMS method has a lack of efficiency because of its convergence $n^{1/3}$.

Least Trimmed Squares (LTS) Regression

It was introduced by Rousseeuw (1984) and is given by

$$\min_{\beta} imize \sum_{i=1}^{h} r_{i:n}{}^{2} \tag{3.1.20}$$

where $r_{i:n}{}^{2} \leq \cdots \leq r_{n:n}{}^{2}$ are the ordered squared residuals and h is to be chosen between $\frac{n}{2}$ and n. The LTS estimators search for the optimal subset of size h which has the least squares fit of the smallest sum of squares residuals. Hence, the LTS estimate of β is then the least square estimate of that subset of size h. For the data comes from continuous distribution breakdown points of LTS equals $\min(n-h+1, h-p+1)/n$, we have $h=[n/2]+[(p+1)/2]$ yields the maximum breakdown point, is asymptotically 50%, whereas $h=n$ gives the ordinary least squares with breakdown point $=1/n$. LTS has the properties such as affine equivariance and asymptotic normality. Its influence function is bounded for both (directions: response and explanatory) the vertical outliers and bad leverage points. Moreover, LTS regression has several advantages over LMS. Its objective function is smoother, making LTS less 'jumpy' (*i.e.*, sensitive to local effects) than LMS. LTS has better statistical efficiency than LMS because of its asymptotically normal property (Hossjer, 1994), whereas LMS has a lower convergence rate (Rousseeuw, 1984). This also makes the LTS more suitable than the LMS as a starting point for two-step estimators such as the MM–estimators (Yohai, 1987) and generalized M-estimators (Simpson, Ruppert and Carrol, 1992; Cookley and Hettmansperger, 1993). It also fails to fit a correct model when large number of clustered outliers exits and with more than 50% outliers in the data. The main drawback of the LTS method is that the objective function requires sorting of the squired residuals, which takes $O(n \log n)$ operations compared with only $O(n)$ operations for the median. The performance of this method has recently been improved by the FAST-LTS (Rousseeuw and van Driessen, 1999) and Fast and robust bootstrap for LTS (Willems and Aelst, 2004).

Reweighted Least Squares (RLS)

The basic principle of LMS and LTS is to fit the majority of the data, after which outliers may be identified as those points that lie far away from the robust fit; that is, the cases

with large positive or large negative residuals. Rousseeuw and Leroy (1987) bring another idea to improve crude LMS and LTS solutions, apply a weighted least squares analysis based on the identification of outliers and use the following weights:

$$w_i = \begin{cases} 1 & if \quad |r_i / \hat{\sigma}| \le 2.5 \\ 0 & if \quad |r_i / \hat{\sigma}| > 2.5 \end{cases} \tag{3.1.21}$$

This means simply that the case i will be retained in the weighted LS if it's LMS residual is small to moderate, but disregarded if it is an outlier. Then they defined weighted least squares by

$$\min imize \quad \sum_{i=1}^{n} w_i r_i^2 . \tag{3.1.22}$$

Therefore, the RLS can be seen as ordinary LS on a 'reduced' data set, consisting of only those observations that received a nonzero weight. The resulting estimator still possesses high breakdown point, but is more efficient in a statistical sense (under Gaussian assumptions).

3.1.4 Reasons for Reluctance to Use Robust Regression

In spite of having some merits of robust regression it has some limitations/reluctances to use as follows:

1. The belief that large sample sizes make robust techniques unnecessary.
2. The belief that outliers can be detected simply by eye or by looking by OLS estimates, or by sensitivity analysis, obviating the need for a robust analysis.
3. Existence of several 'robust regression' techniques with the guidance available as to which is appropriate.
4. Unfamiliarity with interpretation of results from a robust analysis.
5. Unawareness of gains available from robust analysis in real data sets.

3.2 Regression Diagnostics: Deletion Approach

This section gives the various ideas relevant to the regression diagnostics depending upon the deletion approach. Many types of diagnostic measures have been developed bearing different thinking by this time. Measures show that how statisticians try to develop and modify the existing methods time to time. It is evident that none of the method is always gained the success. This is the reason for identifying unusual observations and has seen a grate deal of attention in almost every field of data analysis.

3.2.1 Deletion Approach and Its Main Concepts

The implicit assumption for OLS (Chatterjee and Hadi, 1988) states, 'all the observations are equally reliable and have an equal role for determining the results'. But the reality is different, it is not always possible that all the observation have same role. Belsley *et al.* (1980) pointed out, " the ordinary least squares method may be so ineffective for the estimation of true innovations that any simple type of modification defined for the entire set may not be fruitful for all cases". Deletion approach seems that different observation has different weights on the analysis. There exist a vast number of methods for detecting unusual observations and measuring their effects on various aspects of the analysis. Some are individual and some are combination of interrelated methods that are based on one function of the observations or the structure of the regression model or the function of estimates.

Since the presence of unusual observations affects the estimation, the idea of deleting them (unusual cases) from the analysis and re-estimating the model with the rest has been generally proposed. Case-deletion requires five questions:

- Which should be deleted?
- What should be the criteria for deletion?
- How many cases (observations) should be deleted?
- What should be the role of deleted observations in subsequent analysis?

- How much information may be lost by deletion required number of observation and how much gain can be achieved?

Atkinson (1986) pointed out, "There is no universal agreement among statisticians about all these questions". That is why a number of methods have been suggested and have lot of arguments in favor of the respective methods.

3.2.2 Group Deletion Diagnostic Method

The methods of group (multiple) deletion are as well known as single deletion but the basic difference is, in case of group deletion a number of observations (group) are deleted at a time. Difficulty is identifying the cases to be deleted. Sometimes it is impossible for the large number of observations. For example if consideration is given to deletion of 5 out of 30 cases, there are 1, 42,506 possibilities. This is running with time and space contains. Atkinson (1986) mentioned, in some examples, the sequential employment of single deletion methods leads to the deletion of important sets of observations. He also mentioned that, single deletion diagnostics are affected by masking and swamping (masking occurs when an outlying subset goes undetected because of the presence of another, usually adjacent and swamping occurs, when good observations are incorrectly identified as outliers because of the presence of another, usually remote, subset of observations; Hadi and Simonoff, 1993) phenomena, single deletion diagnostic methods fail to reveal outliers and influential observations. In other words, the importance of the individual observation is not evident unless several observations are not deleted at once.

3.2.3 Deletion of Unusual Observation

There exits a large number of interrelated methods for deleting unusual observations and measuring their effects on various aspects of the analysis. These methods can be divided into seven groups based on the specific aspect of the analysis that one is interested in. Chatterjee and Hadi in 1988 mentioned that, measures may be based on any one of the following quantities:

1. Residuals,
2. Remoteness of points in the X-Y space,
3. Influential curve (center of confidence ellipsoids),
4. Volume of confidence ellipsoids,
5. Likelihood function,
6. Subset of regression coefficients, and
7. Eigen-structure of X.

We have a brief discussion on some of the above quantities that are most popular and relevant to our study.

3.2.4 Measures Based On Residuals

"Residuals play an important role in regression diagnostics; no analysis is complete without a thorough examination of the residuals", Chatterjee and Hadi (1988). To assess the appropriateness of the model

$$Y = X\beta + \varepsilon, \tag{3.2.1}$$

it is necessary to ensure whether the assumptions about the errors are reasonable. The problem here is that the ε_i can neither be observed nor they can be estimated directly. This must be done indirectly using residuals. For the linear least squares, the vector of residual r can be written as

$$\hat{\varepsilon} = r = Y - \hat{Y} = (I - H)\varepsilon \tag{3.2.2}$$

where $H = X(X^T X)^{-1} X^T$

or in scalar form,

$$\hat{\varepsilon}_i = r_i = y_i - \hat{y}_i$$

$$= \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j, \qquad i=1,2,...,n \tag{3.2.3}$$

where h_{ij} is the ij-th element of H. The identity (3.2.2) indicates clearly that the relationship between r and ε depends only on H. If h_{ij}'s are sufficiently small, r will serve

as a reasonable substitute of ε. Furthermore, if the elements of ε are independent and have the same variance, identity (3.2.2) indicates that the ordinary residuals are not independent (unless H is diagonal) and they do not have the same variance (unless the diagonal elements of H are equal), because

$$Var(r_i) = \sigma^2(1 - h_{ii}). \tag{3.2.4}$$

The ordinary residuals are not appropriate for diagnostic purpose; it is preferable to use a transformed version of the ordinary residuals. That is instead of r_i one may use

$$f(r_i, \sigma_i) = \frac{r_i}{\sigma_i}, \tag{3.2.5}$$

where σ_i is the standard deviation of the i-th residual.

Some special types of scaled residuals are given below.

Normalized Residuals

The i-th normalized residual is obtained by replacing σ_i in (3.2.5) by $\sqrt{r^T r}$ as

$$a_i = \frac{r_i}{\sqrt{r^T r}} \qquad i = 1,2,\ldots,n. \tag{3.2.6}$$

Standardized Residuals

To overcome the problem of unequal variances, we standardize the i-th residual r_i by dividing it by its standard deviation (square root of the mean square for error). Residuals have zero mean and approximate average variance is estimated by

$$\frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{n-p} = \frac{\sum_{i=1}^{n} r_i^2}{n-p} = \frac{SS_{\text{Res}}}{n-p} = MS_{\text{Res}} = \frac{r^T r}{n-p} \tag{3.2.7}$$

$$i.e. \quad \hat{\sigma} = \sqrt{\frac{r^T r}{n-p}}.$$

Hence the i-th standardized residual is,

$$d_i = \frac{r_i}{\hat{\sigma}}. \tag{3.2.8}$$

Studentized Residuals

Standardized residual is also referred to an internal scaling of the residual because $\hat{\sigma}^2$ is an internally generated estimate of σ^2 obtained from the fitting of the model to all n observations. The internally Studentized residuals are defined by

$$e_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}. \qquad i=1,2,...,n \tag{3.2.9}$$

Another approach would be to use an estimate of σ^2, based on a data set with the i-th observation deleted. This deleted scaled residual is referred as externally Studentized residual and defined as

$$e_i^* = \frac{y_i - x_i^T \hat{\beta}^{(-i)}}{\hat{\sigma}^{(-i)}\sqrt{(1-h_{ii})}} = \frac{r^{(-i)}}{\hat{\sigma}^{(-i)}\sqrt{1-h_{ii}}} \qquad i=1,2,...,n, \tag{3.2.10}$$

where $\hat{\sigma}^{(-i)2} = \frac{1}{n-p-1}\sum_j (y_j - x_j^T \hat{\beta}^{(-i)})^2$.

After some simplifications we get the relation between the Studentized residuals (external and internal) as

$$e_i^* = e_i \sqrt{\frac{n-p-2}{n-p-1-e_i^2}}. \tag{3.2.11}$$

Under the usual assumptions Ellenberg (1976) showed that externally Studentized residuals follow Student's t_{n-k-1} distribution. Behnken and Draper (1972), Davies and Hutton (1975), and Huber (1975) all of them recommended the external Studentized residuals as more appropriate than the standardized (internal Studentized) residuals for identifying outliers since the effect of i-th observation is more pronounced in the case of the former.

PRESS Residuals

Predictive residual error sum of squares (PRESS) residuals proposed by Allen (1974) as a criterion of model selection. The logic behind the PRESS residual is, if the i-th observation y_i is really unusual, the regression model based on all observations may be

overly influenced by this observation. This could produce a fitted value $\hat{y}_i$ that is very similar to the observed value y_i, and consequently, the ordinary residual r_i will be small. If the i-th observation is deleted, then y_i cannot be influenced by that observation, so the resulting residual should be likely to indicate the presence of the outlier. After deleting the i-th observation if we fit the regression model on the n-1 observations, the corresponding prediction error is

$$r^{(-i)} = y_i - \hat{y}^{(-i)}. \qquad i = 1,2,...,n \tag{3.2.12}$$

Cook and Weisberg (1982) named PRESS residuals as predictive residuals and Atkinson (1985) mentioned it as deletion residuals. It seems that the calculation of PRESS residuals will require fitting different n regressions, but using the results from Miller (1974), it is possible to calculate PRESS residuals from the results of a single least-squares fit to all n observations. The i-th PRESS residual can be written as

$$r^{(-i)} = \frac{r_i}{1-h_{ii}}, \qquad i = 1,2,...,n \tag{3.2.13}$$

Generally, a large difference between the ordinary residual and the predicted residual will indicate a point where the model fits the data well, but a model built without that point predicts poorly. The variance of the i-th PRESS residual is

$$Var(r^{(-i)}) = \frac{\sigma^2}{1-h_{ii}},$$

so that a standardized PRESS residual is

$$\frac{r}{\sqrt{Var(r^{(-i)})}} = \frac{r_i}{\sqrt{\sigma^2(1-h_{ii})}}. \tag{3.2.14}$$

If we use MS_{Res} to estimate σ^2, then it would be the external Studentized residual.

Jackknife Residuals

Tukey (1958) proposed Jackknife method as an extension of the idea of Quenouille (1949), which reduces bias in estimation, and provides approximate confidence intervals in cases where ordinary distribution theory proves difficult. Let us suppose we have a

sample $\mathbf{x}=(x_1,\ldots,x_n)$ and an estimator $\hat{\theta}=f(x)$.The i-th jackknife sample consists the data set with the i-th observation deleted , i.e. , the i-th jackknife sample is defined as

$$\mathrm{x}^{(-i)}=\left(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n\right)^T$$

Let us also define the i-th jackknife replication of $\hat{\theta}$ by

$$\hat{\theta}^{(-i)}=f\left(\mathrm{x}^{(-i)}\right),$$

thus finally the jackknife estimator of θ is defined as

$$\tilde{\theta}=n\hat{\theta}-(n-1)\hat{\theta}_{(.)} \tag{3.2.15}$$

where $$\hat{\theta}_{(.)}=\frac{\sum_{i=1}^{n}\hat{\theta}^{(-i)}}{n}.$$

In a regression problem, the jackknife technique was proposed to obtain less bias in the estimation of the standard errors of regression coefficients, and in constructing precise confidence intervals for them (Miller,1974; Hinkley 1977; Wu 1986).

There are at least two different ways one can define jackknife residuals. Let $\hat{\beta}$ be the usual LS estimator of regression coefficients β and $\hat{\beta}^{(-i)}$ be the corresponding estimates with the i-th case deleted. Then the jackknife estimate of regression coefficients β is by definition

$$\hat{\beta}_{(Jack)}=n\hat{\beta}-\frac{n-1}{n}\sum_{i=1}^{n}\hat{\beta}^{(-i)}. \tag{3.2.16}$$

Using these estimates of regression coefficients one can define the i-th jackknife residual as

$$\tilde{\varepsilon}_i=y_i-x_i^T\hat{\beta}_{(Jack)}. \qquad i=1,2,\ldots..,n \tag{3.2.17}$$

The other way to obtain jackknife residuals is to jackknife the set of LS residuals or some other set of full-sample residuals. The i-th jackknife residual is then defined by

$$\tilde{\varepsilon}_i=n\hat{\varepsilon}_i-\frac{n-1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^{(-i)}. \qquad i=1,2,\ldots,n \tag{3.2.18}$$

Sometimes PRESS residuals are misleadingly referred to as jackknife residuals in the literature (Mosteller and Tukey, 1977).

Delete –*d* Jackknife Residuals

Wu (1986) proposed delete-*d* jackknife technique as a generalization of ordinary jackknife where he left out $d \geq 1$ observations at a time. Let $\hat{\theta}^{(-D)}$ denote $\hat{\theta}$ applied to the data set after the deletion of a group D of size $d \geq 1$. Then the delete-*d* jackknife estimator of θ is

$$\hat{\theta}_{(Jack)}^{(-D)} = n\hat{\theta} - (n-d)\hat{\theta}_{(.)} \tag{3.2.19}$$

where

$$\hat{\theta}_{(.)} = \frac{\sum_R \hat{\theta}^{(-D)}}{\binom{n}{d}},$$

and the sum is over all subset of R of size *(n-d)* chosen without replacement from the data set. Like the ordinary jackknife there are two ways of calculating deleted-*d* jackknife residuals. The standard practice is to use this method to estimate the regression parameters and the residuals are estimated afterwards. The other way to calculating them is to apply delete-*d* jackknife technique on LS residuals.

Let R be a subset of 'remaining' in the analysis after the 'deletion' of a group of d observations, indexed by D, from the data set. Without loss of generality, assume that the deleted cases are last d rows of X, Y, and ε. Let $\hat{\beta}^{(-D)}$ be the LS estimate of β based on the $\{y_i, x_i\}$ in R, that is

$$\hat{\beta}^{(-D)} = \left(X_R^T X_R\right)^{-1} X_R^T Y_R \tag{3.2.20}$$

where $Y_R = (y_1, \cdots, y_{n-d})^T$ and $X_R = \left[x_{i1}, \cdots, x_{i(n-d)}\right]^T$.

Wu (1986) defined a residual vector of order *(n-d)* computed on subset R by

$$\hat{\varepsilon}_R = Y_R - X_R\hat{\beta} \tag{3.2.21}$$

and the delete-*d* jackknife technique can be applied to them.

It is, however, possible to obtain a full set of residuals by the use of deleted-d jackknife methodology. Again there are at least two ways of computing a full set of delete-d jackknife residuals. The first category of delete-d jackknife residuals can be defined by

$$\hat{\varepsilon}_{(Jack)}^{(-D)} = Y - X\hat{\beta}_{(Jack)}^{(-D)}, \tag{3.2.22}$$

where $$\hat{\beta}_{(Jack)}^{(-D)} = n\hat{\beta} - \frac{n-d}{\binom{n}{d}}\sum_{R}\hat{\beta}^{(-D)}$$

The other type of delete-d jackknife residuals are

$$\tilde{\varepsilon} = n\hat{\varepsilon}_i - \frac{n-d}{\binom{n}{d}}\sum_{R}\hat{\varepsilon}_R^{(-D)} \qquad i = 1,2,...,n \tag{3.2.23}$$

where $\hat{\varepsilon}_R^{(-D)} = Y_R - X_R\hat{\beta}^{(-D)}$, is the set of LS residuals in a subset indexed by R.

Bootstrap Residuals

Efron's (1979) first article on bootstrap technique synthesized some of the earlier resampling ideas and established a new framework for simulation based statistical analysis. The idea of replacing complicated and often inaccurate approximations to biases, variances, and other measures of uncertainty by computer simulations caught the imagination of both theoretical researchers and users of statistical methods. Efron and Tibshirani (1993), and Shao and Tu (1995) presented excellent reviews of all these studies. There are two different ways in which one could generate a distribution of bootstrap residuals. Firstly, one could estimate a set of residuals by the LS method and then can generate bootstrap residuals by random sub sampling from that set. In other words, a distribution of bootstrap residuals can be obtained by bootstrapping LS residuals. The other approach is by bootstrapping cases to fit the regression model at first and then use it to estimate residuals. There are again two different approaches of bootstrapping a regression model. One is based on drawing an *i.i.d.* sample $\{\varepsilon_i^*\}$ of size n from the LS residuals $\{\hat{\varepsilon}_i\}$.

The bootstrap observations are then defined by

$$y_i^* = x_i^T \hat{\beta} + \varepsilon_i^*, \qquad i = 1,2,...,n \tag{3.1.24}$$

by treating $\hat{\beta}$ as the *true parameter* and $\{\hat{\varepsilon}_i\}$ as the *population* of innovations.

Thus the Bootstrap Least Square (BLS) estimate of β for each bootstrap sample is obtained by

$$\beta^* = \left(X^T X\right)^{-1} X^T Y^* \tag{3.2.25}$$

where $Y^* = \left(y_1^*, \cdots, y_n^*\right)^T$.

The second bootstrap method is based on drawing n bootstrap samples of pairs $\{(y_i^*, x_i^*)\}$ from the n observed pairs $\{(y_i, x_i)\}$.

Then the BLS estimate of β is computed as

$$\beta^* = \left(\sum_{i=1}^{n} x_i^* x_i^{*T}\right)^{-1} \sum_{i=1}^{n} x_i^* y_i^*. \tag{3.2.26}$$

Efron and Gong (1983) advocated use of the second bootstrap method for its lower sensitivity to assumptions. Whatever the choice of β^*, the overall bootstrap estimate of regression coefficients is then obtained by

$$\hat{\beta}_{BOOT} = \frac{1}{N} \sum_{i=1}^{N} \beta_i^* \qquad i=1,2..., N \tag{3.2.27}$$

where N is the number of bootstrap replications.

One can now define a full set of bootstrap residuals as the differences between observed and fitted responses when regression coefficients are estimated by $\hat{\beta}_{BOOT}$, computed by either of these two approaches.

3.2.5 Measures Based on Remoteness of Points in *X-Y* space (Distance Measures)

Chatterjee and Hadi (1988) pointed, "Examination of residuals alone is not sufficient for detecting unusual observations, especially those corresponding to high-leverage points. This can be seen from the property that, $0 \le \left(h_{ii} + r_i^2 / r^T r\right) \le 1$, which implies that

observations with large h_{ii} tend to have small residual and therefore go undetected in the usual plots of residuals".

In this sub section we will discuss some related quantities for measuring the leverage of a point.

Diagonal Elements of Hat Matrix, H

We know that the diagonal elements of the hat (prediction) matrix, $h_{ii} = x_i^T \left(X^T X\right)^{-1} x_i$, play an important role in determining the fitted values, the magnitude of the residuals, and their variance-covariance structure. For these reasons, Hoaglin and Welsch in 1978 suggested the examination of both e_i^* and $h_{ii}(e_i^*$ for detecting outliers and h_{ii} for detecting high-leverage points that are potentially influential), and added, "These two aspects of the search for troublesome data points are complementary; neither is sufficient by itself". According to Hocking and Pendleton (1983), high-leverage points,...are those for which the input vector x_i is, in some sense, far from the rest of the data'. But the question here is "how far is far?" some common suggested cut-off points for h_{ii} are as follows:

(a) The reciprocal of h_{ii} can be thought of as the effective or equivalent number of observations that determine $\hat{y}_i$ (Huber, 1977 and 1981). Huber (1981) suggested that points with

$$h_{ii} \geq 0.2 \tag{3.2.28}$$

be classified as high-leverage points.

(b) Hoaglin and Welsch (1978) suggested that points with

$$h_{ii} \geq \frac{2p}{n} \tag{3.2.29}$$

and Vellman and Welsch (1981) suggested that points with

$$h_{ii} \geq \frac{3p}{n} \tag{3.2.30}$$

be classified as high-leverage points.

(c) If a regression model contains a constant term and the rows of X are *i.i.d.* $N_p(\mu,\Sigma)$,

then
$$\frac{n-p-1}{p}\frac{h_{ii}-\frac{1}{n}}{1-h_{ii}} \approx F_{(p,n-p-1)}, \tag{3.2.31}$$

which lead to nominating points with

$$h_{ii} \geq \frac{nF(p)+(n-p-1)}{nF(p)+n(n-p-1)}, \tag{3.2.32}$$

as high-leverage points, where F is the *100(1-α)* point of $F_{(p,n-p-1)}$.

The suggested cut-off points for h_{ii} should not be used mechanically; they should serve as rough yardsticks for general guidance and flagging of troublesome points. This is best accomplished by graphical display of h_{ii} such as index plot, stem-leaf display, and/or box plots.

Mahalanobis Distance

Distance measures are useful for identifying high-leverage points. The Mahalanobis distance (1936) introduced by himself is the first and most well known distance measure in multivariate analysis and is introduced as

$$MD_i = \sqrt{(\mathrm{x_i}-\mathrm{T(X)})\,\mathrm{C(X)}^{-1}(\mathrm{x_i}-\mathrm{T(X)})^{\mathrm{T}}} \tag{3.2.33}$$

where, T(X) and C(X) are the sample mean and covariance matrix and defined as

$$\mathrm{T(X)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{x_i} \tag{3.2.34}$$

and
$$\mathrm{C(X)} = \frac{1}{\mathrm{n-1}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i}-\mathrm{T(X)})(\mathrm{x_i}-\mathrm{T(X)}^{T}). \tag{3.2.35}$$

This measure tells us how far $\mathbf{x}_i$ is from the center of the cloud, taking into account the shape of the cloud as well. Suppose that X contains a column of ones and $\tilde{X}$ denote centered X excluding the constant column. A statistic which measure how far $\mathrm{x_i}$ is from the center of the data set is commonly computed as $(n-1)^{-1}\,\tilde{\mathrm{x}}_i^{\mathrm{T}}(\tilde{\mathrm{X}}^{\mathrm{T}}\tilde{\mathrm{X}})^{-1}\,\tilde{\mathrm{x}}_i$, where $\tilde{\mathbf{x}}_i$ is the i-th row of $\tilde{X}$. We are interested in measuring how far $\mathrm{x_i}$ is from the rest of the other

observations; hence it is natural to exclude x_i when computing the mean and variance of X. Therefore Mahalanobis distance can be written as

$$MD_i = (n-2)\{\tilde{x}_i - \bar{\tilde{X}}_{(i)}\}^T \{\tilde{X}_{(i)}^T (I-(n-1)^{-1}11^T)\tilde{X}_{(i)})\}^{-1}\{\tilde{x}_{(i)} - \bar{\tilde{X}}_{(i)}\} \qquad (3.2.36)$$

where, $\bar{\tilde{X}}_{(i)}$ is the average of $\tilde{X}_{(i)}$ and $\bar{\tilde{X}}_{(i)} = (n-1)^{-1}\tilde{X}_{(i)}^T 1 = -(n-1)^{-1}\tilde{x}_i$.

Finally we get the distance as

$$MD_i = \frac{n(n-2)}{n-1}\frac{h_{ii} - 1/n}{1-h_{ii}} \qquad i = 1,2,...,n. \qquad (3.2.37)$$

Mahalanobis distance suffers from the masking effect, by which multiple outliers do not necessarily have a large MD_i. This can be realized from the equivalence relation:

$$h_{ii} = \frac{(MD_i)^2}{n-1} + \frac{1}{n}. \qquad (3.2.38)$$

Minimum Volume Ellipsoid

The minimum volume ellipsoid (MVE) estimator is defined as the pair *(T, C)*, where *T(X)* is a *p*-vector and *C(X)* is a positive semi-definite $p \times p$ matrix such that the determinant of *C* is minimized subject to

$$\{i; \quad (x_i - T)C^{-1}(x_i - T)^T \le a^2\} \ge h, \qquad (3.2.39)$$

where $h = [(n+p+1)/2]$. The number a^2 is a fixed constant, which can be chosen as $\chi^2_{p,.5}$ when we expect majority of the data to come from a normal distribution. For small samples one also needs a factor $c^2_{n,p}$, which depends on *n* and *p*. The MVE has a breakdown point of nearly 50%.

Robust Distance

Mahalanobis distance is not robust, because *T(X)* and *C(X)* are not robust. Therefore it seems natural to replace *T(X)* and *C(X)* by robust estimators. The first such estimator was proposed by Stahel (1981) and Donoho (1982). Rousseeuw and van Zomeren (1990) proposed robust distance RD_i by inserting the MVE (minimum volume ellipsoid)

(Rousseeuw, 1985) estimates for *T(X)* and *C(X)* as an outliers and high-leverage points diagnostic measure.

Minimum Covariance Determinant

The minimum covariance determinant (MCD) estimator is another method with high break down point (Rousseeuw, 1985). It searches for a subset containing half of the data, the covariance matrix of which has the smallest determinant. It has been proved by Butler *et al.,* (1993) that MCD estimator is asymptotically normal but it needs somewhat more computation time than MVE.

Weighted Squared Standardized Distance

Suppose that model (3.2.1) contains a constant term and define

$$c_{ij} = \hat{\beta}_j\left(x_{ij} - \overline{X}_j\right), i = 1,2,...,n \quad and \quad j = 1,2,...,k \quad , \tag{3.2.40}$$

where $\overline{X}_j$ is the average of the j-th column of X. the quantity c_{ij}, $i=1,2,..,n,$ and $j=1,2,..,k$ may be regarded as the effect of the j-th variable on the i-th predicted value $\hat{y}_i$. It is easy to show that

$$\hat{y}_i - \overline{Y} = \sum_{j=1}^{k} c_{ij} \ , \tag{3.2.41}$$

where $\overline{Y}$ is the average of Y. Daniel and Wood (1971) suggested Weighted Squared Standardizes Distance, namely,

$$WSSD_i = \frac{\sum_{j=1}^{k} c_{ij}^2}{s_Y^2}, \qquad i=1,2,..,n, \tag{3.2.42}$$

where $$s_Y^2 = \frac{\sum_{j=1}^{n}\left(y_j - \overline{Y}\right)^2}{n-1},$$

to measure the influence of the i-th observation on moving the prediction at the i-th point from $\overline{Y}$. Thus $WSSD_i$ is a measure of the sum of squared distances of x_{ij} from the average of the j-th variable, $\overline{X}_j$, weighted by the relative importance of the j-th variable (the magnitude of the estimated regression coefficients). Therefore $WSSD_i$ will be large if

the i-th observation is extreme in at least one variable whose estimated regression coefficient is large in magnitude.

Diagonal Elements of H_Z (Z, augmented matrix)

A diagonal element from prediction matrix ignores the information contained in Y. It can be augmented the matrix X by the vector Y, and obtain $Z=(X: Y)$. Now the diagonal elements h_{zii} of the corresponding prediction matrix H_Z can be written as

$$h_{zii} = z_i^T (Z^T Z)^{-1} z_i = h_{ii} + r_i^2 / r^T r \tag{3.2.43}$$

and thus h_{zii} will be large whenever h_{ii} is large, r_i^2 is large, or both. Hence h_{zii} cannot distinguish between the two distinct situations, a high-leverage point in the X space and an outlier in the Z space.

Hadi's Influence Measure (Potential)

Hadi (1992) mentioned in his paper that in the presence of a high-leverage point the information matrix may breakdown and hence the observations may not have the appropriate leverages. He introduced a single case deletion measure of leverage named by potentials and defined as

$$p_{ii} = x_i^T (X^{(-i)T} X^{(-i)})^{-1} x_i, \tag{3.2.44}$$

where $X^{(-i)}$ is the matrix with i-th observation deleted.

Generalized Potentials

Imon (1996) introduced generalized potentials by using group deletion for all the observations in a data set as

$$p_{ii}^{(-D)} = \begin{cases} \dfrac{p_{ii}^{(-D)}}{1 - p_{ii}^{(-D)}} \; for \quad i \in R \\ p_{ii}^{(-D)} \; for \quad i \in D \end{cases} \tag{3.2.45}$$

where D is the deletion group and R is the remaining set of observations.

3.2.6 Measures Based on the Influence Curve

Measures of the influence of the i-th observation on the regression results are introduced by Hampel (1968, 1974). Some very popular measures that are derived from that idea are given below.

Cook's Distance

Cook (1977,1979) has suggested a way for reducing the influence curve, using a measure of the squared distance between the least-squares estimate based on all n points $\hat{\beta}$ and the estimate obtained by deleting the i-th point, say $\hat{\beta}^{(-i)}$. This distance measure can be expressed in general form as

$$CD_i(M,c)=\frac{\left(\hat{\beta}^{(-i)}-\hat{\beta}\right)^T M\left(\hat{\beta}^{(-i)}-\hat{\beta}\right)}{c}, \qquad i=1,2,.....,n \quad (3.2.46)$$

The usual choice of M and c are $M=X^TX$ and $c=pMS_{Res}$, so that equation (3.2.46) becomes

$$CD_i\left(X^TX,pMS_{Res}\right)\equiv CD_i=\frac{\left(\hat{\beta}^{(-i)}-\hat{\beta}\right)^T X^TX\left(\hat{\beta}^{(-i)}-\hat{\beta}\right)}{pMS_{Res}}, \quad i=1,2,...,n \quad (3.2.47)$$

points with the values of CD_i have considerable influence on the least-squares estimate $\hat{\beta}$. The magnitude of CD_i is usually assessed by comparing it to $F_{\alpha,p,n-p}$. If $CD_i=F_{0.5,p,n-p}$, then deleting point i would move $\hat{\beta}^{(-i)}$ to the boundary of an approximate 50% confidence region for β based on the complete data set. This is a large displacement and indices that the least-squares estimate is sensitive to the i-th data point. Since $F_{0.5,p,n-p}\approx 1$, Cook and Weisberg (1982) suggested considering points to be influential for which $CD_i>1$. Beckman and Trussell (1974) showed that

$$\hat{\beta}-\hat{\beta}^{(-i)}=(X^{(-i)T}X^{(-i)})^{-1}x_i[Y_i-x_i^T\hat{\beta}], \quad (3.2.48)$$

and according to Bingham (1977), CD_i can also be written as

$$CD_i=\frac{(\hat{Y}-\hat{Y}^{(-i)})^T(\hat{Y}-\hat{Y}^{(-i)})}{k\sigma^2}. \qquad i=1,2,...,n \quad (3.2.49)$$

DFBETAS

Belsley *et al.* (1980) introduced measures of deletion influence. The first of these is a statistic that indicates how much the regression coefficient $\hat{\beta}_j$ changes, in standard deviation units, if the *i*-th observation is deleted. This statistic is

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_j^{(-i)}}{\sqrt{S^{(-i)2} C_{jj}}}, \tag{3.2.50}$$

where C_{jj} is the *j*-th diagonal element of $\left(X^T X\right)^{-1}$ and $\hat{\beta}_j^{(-i)}$ is the *j*-th regression coefficient computed without use of the *i*-th observation. Belsley *et al.* (1980) suggested a cutoff of $2/\sqrt{n}$ for $DFBETAS_{j,i}$; that is, if $\left|DFBETAS_{j,i}\right| > 2/\sqrt{n}$, and then the *i*-th observation warrants examinations.

Welsch-Kuh Distance (DFFITS)

The influence of the *i*-th observation on the predicted value can be measured by the change in the prediction at x_i when the *i*-th observation is omitted, relative to the standard error of, that is,

$$\frac{\left|\hat{y} - \hat{y}_i^{(-i)}\right|}{\sigma\sqrt{h_{ii}}} = \frac{\left|x_i^T(\hat{\beta} - \hat{\beta}^{(-i)})\right|}{\sigma\sqrt{h_{ii}}}. \tag{3.2.51}$$

The denominator is just a standardization, since $Var\left(\hat{y}_i\right) = \sigma^2 h_{ii}$. Welsh and Kuh (1977) and Welsch and Peters (1978) suggested using $\hat{\sigma}^{(-i)}$ as an estimate of σ in (3.2.51). Belsley *et al.* (1980) called Welsch-Kuh's distance (*WK*) as $DFFITS_i$, because it is the scaled difference between $\hat{y}$ *and* $\hat{y}^{(-i)}$. Belsley *et al.* (1980) suggests that any observation for which $\left|DFFITS_i\right| > 2/\sqrt{n}$ warrants attention.

Welsch's Distance

Using the empirical influence curve based on *(n-1)* observations, as an approximation to the influence curve for $\hat{\beta}$ and setting

$$M = X^{(-i)T} X^{(-i)} = \left(X^T X - x_i x_i^T\right) \text{ and } c = (n-1)\hat{\sigma}^{(-i)2},$$

Welsch's distance defined as

$$W_i^2 = CD_i\left\{X_{(i)}^T X_{(i)}, (n-1)\hat{\sigma}_{(i)}^2\right\}. \tag{3.2.52}$$

Welsch's distance can be defined by Welsch-Kuh's distance as

$$W_i = WK_i\sqrt{\frac{n-1}{1-p_{ii}}}. \tag{3.2.53}$$

Welsch (1982) has suggested using W_i as a diagnostic tool. The fact that WK_i is easier to interpret has led some authors to prefer WK_i over W_i. It is clear from (3.1.53), however, that W_i gives more emphasis to high-leverage points. Equation (3.2.53) suggests that the cut-off points for W_i can be obtained by multiplying the cut-off points for WK_i by $\{n(n-1)/(n-p-1)\}^{1/2}$. Thus for example, if $2\sqrt{(p+1)/(n-p-1)}$ is used as cut-off point for WK_i, then the corresponding cut-off point for W_i would be $(2/(n-p-1))\sqrt{(p+1)n(n-1)}$. However, if n is large as compared to p, this quantity is approximately $3\sqrt{p+1}$.

Modified Cook's Distance

Atkinson (1981) has suggested using a modified version of Cook's distance for the detection of influential observations. The suggested modification involves replacing $\hat{\sigma}^2$ by $\hat{\sigma}^{(-i)2}$, taking the square root of C_i, and adjusting C_i for the sample size. This modification of the C_i yields

$$C_i^* = \sqrt{CD_i\left(X^T X, \frac{(p+1)(n-1)^2}{n-p-1}\hat{\sigma}^{(-i)2}\right)} \tag{3.2.54}$$

$$= |e_i^*|\sqrt{\frac{h_{ii}}{1-h_{ii}}\frac{n-p-1}{p+1}} \tag{3.2.55}$$

$$= WK_i\sqrt{\frac{n-p-1}{p+1}}. \tag{3.2.56}$$

Atkinson (1981) claims that this modification improves C_i in three ways, namely,

(a) C_i^* gives more emphasis to extreme points,

(b) C_i^* becomes more suitable for graphical displays such as normal probability points, and

(c) for the perfectly balanced case where $h_{ii} = (p+1)/n$, for all i, the plot of C_i^* is identical to that of $|e_i^*|$.

The essential difference among C_i, WK_i, W_i, and C_i^* is in the choice of scale. C_i measures the influence of the i-th observation on $\hat{\beta}$ only, whereas WK_i, W_i, and C_i^* measure the influence on both $\hat{\beta}$ and $\hat{\sigma}^2$.

Table 3.1 The influence measures D_i (M, c) for several choices of of M and c

M	c	Measure
$X^T X$	$(n-1)^2 k\hat{\sigma}^2$	$C_i = \frac{1}{k}\frac{h_{ii}}{1-h_{ii}} r_i^2$
$X^T X$	$(n-1)\hat{\sigma}^{(-i)2}$	$WK_i = \lvert e_i^* \rvert \sqrt{\frac{h_{ii}}{1-h_{ii}}}$
$X^T X$	$n\hat{\sigma}^{(-i)2}$	$WK_i = \lvert e_i^* \rvert \sqrt{\frac{h_{ii}}{1-h_{ii}}}$
$X^T X$	$\frac{(n-1)^2 k}{n-k}\hat{\sigma}^{(-i)2}$	$C_i^* = \lvert e_i^* \rvert \sqrt{\frac{h_{ii}}{1-h_{ii}}\frac{n-k}{k}}$
$X^{(-i)T} X^{(-i)}$	$(n-1)\hat{\sigma}^{(-i)2}$	$W_I^2 = (n-1) e_i^{*2} \frac{h_{ii}}{(1-h_{ii})^2}$

3.2.7 Measures Based on Volume of Confidence Ellipsoids

The diagnostic measures based on the influence curve can be interpreted as given above, as measures which are based on the on the change in the center of the confidence ellipsoid when the *i*-th observation is deleted. An alternative class of measures of influence of *i*-th observation is based on the change in the volume of the confidence ellipsoid when *i*-th observation is deleted. Here are two most mentionable are given below.

Andrews-Pregibon Statistic

The volume of joint confidence ellipsoid for β is inversely proportional to the square root of $\det(X^T X)$. Hence an important criterion in the theory of optimal experimental design is based on $\det(X^T X)$, because large values of $\det(X^T X)$ are indicative to informative designs. Thus the influence of the *i*-th observation on the volume of confidence ellipsoids can be measured by comparing $\det(X^T X)$ and $\det(X^{(-i)T} X^{(-i)})$. On the other hand, omitting an observation with a large residual will result in a large reduction in the residual sum of squares, SSE. The influence of the *i*-th observation can be measured by combining these two ideas and computing the change in both $r^T r$ and $\det(X^T X)$ when the *i*-th observation is omitted.

Andrews and Pregibon (1978) suggested the ratio

$$\frac{SSE^{(-i)} \det(X^{(-i)T} X^{(-i)})}{SSE \det(X^T X)}, \qquad i=1,2,....,n \qquad (3.2.57)$$

Define $Z=(X:Y)$ and thus (3.2.57) becomes

$$\frac{SSE^{(-i)} \det(X^{(-i)T} X^{(-i)})}{SSE \det(X^T X)} = \frac{\det(Z^{(-i)T} Z^{(-i)})}{\det(Z^T Z)} \qquad (3.2.58)$$

which measures the relative change in $\det(Z^T Z)$ due to the omission of the *i*th observation. Omitting an observation that is far from the center of the data will result in a large reduction in the determinant and thereby large increase in the volume. Hence, small values of (3.2.58) call for special attention. For convenience we define

$$AP_i = 1 - \frac{\det\left(Z^{(-i)T} Z^{(-i)}\right)}{\det\left(Z^T Z\right)}. \tag{3.2.59}$$

Cook-Weisberg Statistic

Under normality, the 100(1-α) % joint confidence ellipsoid for β is

$$E = \left\{ \beta : \frac{\left(\beta - \hat{\beta}\right)^T \left(X^T X\right)\left(\beta - \hat{\beta}\right)}{(p+1)\hat{\sigma}^2} \le F_{(\alpha;p+1,n-p-1)} \right\}, \tag{3.2.60}$$

when the i-th observation is omitted, (3.2.60) becomes

$$E^{(-i)} = \left\{ \beta : \frac{\left(\beta - \hat{\beta}^{(-i)}\right)^T \left(X^{(-i)T} X^{(-i)}\right)\left(\beta - \hat{\beta}_{(i)}\right)}{(p+1)\hat{\sigma}^{(-i)2}} \le F_{(\alpha;p+1,n-p-2)} \right\}. \tag{3.2.61}$$

Cook and Weisberg (1980) proposed the logarithm of the ratio of E to $E^{(-i)}$ as a measure of the influence of the i-th observation on the volume of confidence ellipsoid for β, namely,

$$CW_i = \log \frac{Volume(E)}{Volume\left(E^{(-i)}\right)}. \qquad i = 1,2,...,n \tag{3.2.62}$$

Since the volume of an ellipsoid is proportional to the inverse of the square root of the determinant of the associated matrix of the quadratic form, (3.2.62) reduces to

$$CW_i = \log\left\{ \left(\frac{\det\left(X^{(-i)T} X^{(-i)}\right)}{\det\left(X^T X\right)} \right)^{1/2} \left(\frac{\hat{\sigma}}{\hat{\sigma}^{(-i)}} \right)^k \left(\frac{F_{(\alpha;p+1,n-p-1)}}{F_{(\alpha;p+1,n-p-2)}} \right)^{(p+1)/2} \right\}. \tag{3.2.63}$$

Substituting

$$\hat{\sigma}^{(-i)2} = \hat{\sigma}^2 \left(\frac{n-p-1-e_i^2}{n-p-2}\right); \text{ where } e_i = \frac{r_i}{\hat{\sigma}\sqrt{(1-h_{ii})}}$$

and $\det(X^{(-i)T} X^T) = \det(X^T X)(1-h_{ii})$ in (3.2.63), we obtain

$$CW_i = \frac{1}{2}\log(1-h_{ii}) + \frac{k}{2}\log\left(\frac{n-p-2}{n-p-1-r_i^2}\right) + \frac{k}{2}\log\left(\frac{F_{(\alpha;p+1,n-p-1)}}{F_{(\alpha;p+1,n-p-2)}}\right). \tag{3.2.64}$$

Cook and Weisberg (1980) say this about CW_i: "if this quantity is large and positive, then deletion of the i-th case will result in a substantial decrease in volume...[and if it is] large and negative, the case will result in a substantial increase in volume." Inspection of (3.2.64) indicates that CW_i will be large and negative where e_i^2 is small and h_{ii} is large, and it will be large and positive where e_i^2 is large and h_{ii} is small. But if e_i^2 and h_{ii} are either large or small, then CW_i tends toward zero.

3.2.8 Measures Based on a Subset of the Regression Coefficients

The influence measures discussed thus far assume that all regression coefficients are of equal interest. Diagnostic measures that involve all regression coefficients may some times be non informative and misleading (Comments Pregibon, in Atkinson 1982). It may happen that an observation is influential only on one dimension (variable). Also an observation with a moderate influence on all regression coefficients may be judged more influential than one with a large influence on one coefficient and a negligible influence on all others. Information about the influence of an observation on a subset of the regression coefficient is, therefore, of interest.

3.2.9 Measures Based on Eigen-structure of X

It is known that the eigenstructure of X can change substantially when a row is added to or omitted from X. Statisticians study the influence of the i-th row of X on the eigenstructure of X in general and on its condition number and collinearity indices in particular. Except for very special cases, no closed form expression connecting the eigenstructure of X to that of $X^{(-i)}$ exists. Some concepts in numerical analysis are used to define the condition number and collinearity indices of a given matrix, with graphical illustrations that individual or small groups of observations can have substantial influence on these measures of collinearity.

3.2.10 Limitations of Deletion Approach

Deletion diagnostic approach has some limitations like any other diagnostic approach, some are as follows:

(1) Single deletion diagnostics are affected by masking and swamping phenomena.

(2) Multiple deletion diagnostic methods are sometimes impossible due to combinatorial problem.

(3) Identification of suspect cases that have to be deleted is a difficult task. Especially group deletion diagnostic methods heavily depend on robust regression/or single deletion diagnostic methods.

(4) Limitations of prior identification methods can make the group deletion results erroneous.

(5) Diagnostic techniques that are free from robustness can make the decisions non robust and may be misleading.

Chapter 4

"For the things we have to learn before we can do them,
we learn by doing them."

Aristotle

Identification of Influential Observations in Linear Regression

The usual way to measure the influence of an observation in a regression analysis is to delete the observation from the data set and compute a convenient norm of the change in the parameters or in the vector of forecasts. In this chapter we propose two new measures based on deletion approach (single and group) for identifying influential observations in simple and multiple linear regressions. We also study the different aspects of the measures. We demonstrate the calculation and show the advantages of the proposed measures in the identification of simple and multiple influential cases through several well-referred data sets.

4.1 Squared Difference in Beta (SDFBETA)

We introduce a possible diagnostic measure for measuring the influence of the i-th (single) observation. The measure is originated from the idea of Cook's distance (1977) and brings a modification in Difference in Beta (DFBETA) in Belsely *et al.* (1980). We name this measure Squared Difference in Beta (SDFBETA). The suggested measure is defined to be

$$SDFBETA_i = \frac{(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)^T (X^T X)(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)}{s^{(-i)^2}(1 - h_{ii})} \tag{4.1}$$

where $\hat{\beta}_i^{(-i)}$ is the estimated i-th parameter and $s^{(-i)^2}(1-h_{ii})$ is the variance of the i-th residual respectively after deleting i-th observation.

4.1.1 Relation with Cook's Distance and DFFITS

We see in Belsely *et al.* (1980) that DFFIT, DFFITS, DFBETA and Cook's Distance (CD) can be defined as

$$DFFIT_i = \hat{y}_i - \hat{y}_i^{(-i)} = x_i^T(\hat{\beta}_i - \hat{\beta}_i^{(-i)}) = \frac{h_{ii}\hat{\varepsilon}_i}{1-h_{ii}}, \tag{4.2}$$

where $\hat{\varepsilon}_i = y_i - \hat{y}_i$ and h_{ii} is the i-th diagonal element of the leverage matrix.

We have, $$DFFITS_i = \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{\sigma^{(-i)}\sqrt{h_{ii}}} = \left[\frac{h_{ii}}{1-h_{ii}}\right]^{1/2} \frac{\hat{\varepsilon}_i}{s^{(-i)^2}\sqrt{(1-h_{ii})}} \tag{4.3}$$

$$DFBETA_i = \hat{\beta}_i - \hat{\beta}_i^{(-i)} \frac{(X^TX)^{-1}x_i^T\hat{\varepsilon}_i}{1-h_{ii}} \tag{4.4}$$

and $$CD_i = \frac{(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)^T(X^TX)(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)}{ps^2}. \tag{4.5}$$

After simplification (see section 2.6), Cook's distance can be defined as

$$CD_i = \frac{\hat{\varepsilon}_i^2 h_{ii}}{ps_i^2(1-h_{ii})^2} \tag{4.6}$$

$$= \frac{h_{ii}}{1-h_{ii}} \cdot \frac{\hat{\varepsilon}_i^2}{s^{(-i)^2}(1-h_{ii})} \cdot \frac{s^{(-i)^2}}{ps_i^2}$$

$$= (DFFITS_i)^2 \cdot \frac{s^{(-i)^2}}{ps_i^2} \tag{4.7}$$

and $$DFFITS_i^2 = \frac{CD_i ps_i^2}{s^{(-i)^2}}, \tag{4.8}$$

hence $$SDFBETA_i = \frac{(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)^T(X^TX)(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)}{s^{(-i)^2}(1-h_{ii})} \tag{4.9}$$

$$= \frac{CD_i ps_i^2}{s^{(-i)^2}(1-h_{ii})} \tag{4.10}$$

$$= \frac{DFFITS_i^2}{1-h_{ii}}. \tag{4.11}$$

4.1.2 Cut-off Value for SDFBETA

It may not be easy to find a theoretical distribution of *SDFBETA*, however it should not make any problem to get a confidence bound type cut-off value for them. We use the equation (4.11), $|DFFITS| \geq 3\sqrt{k/n}$ (like Belsely *et al.* 1980) and $h_{ii} \geq 3p/n$ (Vellman and Welsch, 1981) to make the cut off value for SDFBETA. As a result, we may consider the *i*-th observation to be influential if it satisfies the condition

$$\text{SDFBETA}_i \geq \frac{(3\sqrt{k/n})^2}{1-(3p/n)} = \frac{9k}{n-3p},\ k = p+1. \tag{4.12}$$

4.1.3 Examples

The measure, squared difference in beta (SDFBETA), is a single case deletion measure. At first we show its performance for simple influential case identification and then for identification of multiple influential observations.

Monthly Payments Data

The following real data is extracted from Rousseeuw and Leroy (1987), Rousseeuw *et al.* (1984); the data shows monthly payments of large Belgian insurance company in1979, made as a result of the end of period of life-insurance contracts. In December a very large sum was paid, because of one extremely high supplementary pension. We see simply from the scatter plot (Figure 4.1 (a)) of the data set December situation is not usual. Table 4.1 presents few single case diagnostics to identify the influential observations. The cut-off values for each of the statistics are given inside the braces. Sometimes standardized residuals are used to detect outlier and cases having values grater than 2.5 are suspects.

For this data only observation 12 (December) is under suspect. We see from figure 4.1 (b,c, and d) that *SDFBETA* successfully treats case December as an influential as well as Cook's distance and *DFFITS* and at the same time it confirms the presence of single influential observation.

Table 4.1 Single case diagnostics for Monthly Payments data

Month	Payment	st.res (2.50)	h_{ii} (0.25)	CD_i (1.000)	$\|DFFITS_i\|$ (0.82)	$\|SDFBETA_i\|$ (2.000)
1	3.22	0.248	0.295	0.013	0.152	0.033
2	9.62	0.784	0.225	0.089	0.413	0.221
3	4.5	0.085	0.169	0.001	0.036	0.002
4	4.94	-0.008	0.127	0.000	-0.003	0.000
5	4.02	-0.233	0.099	0.003	-0.073	0.006
6	4.2	-0.345	0.085	0.006	-0.100	0.011
7	11.24	0.223	0.085	0.002	0.065	0.005
8	4.53	-0.580	0.099	0.019	-0.186	0.038
9	3.05	-0.875	0.127	0.056	-0.330	0.124
10	3.76	-0.961	0.169	0.094	-0.432	0.224
11	4.23	-1.088	0.225	0.172	-0.592	0.452
12	**42.69**	**3.063**	**0.295**	**1.961**	**7.550**	**80.842**

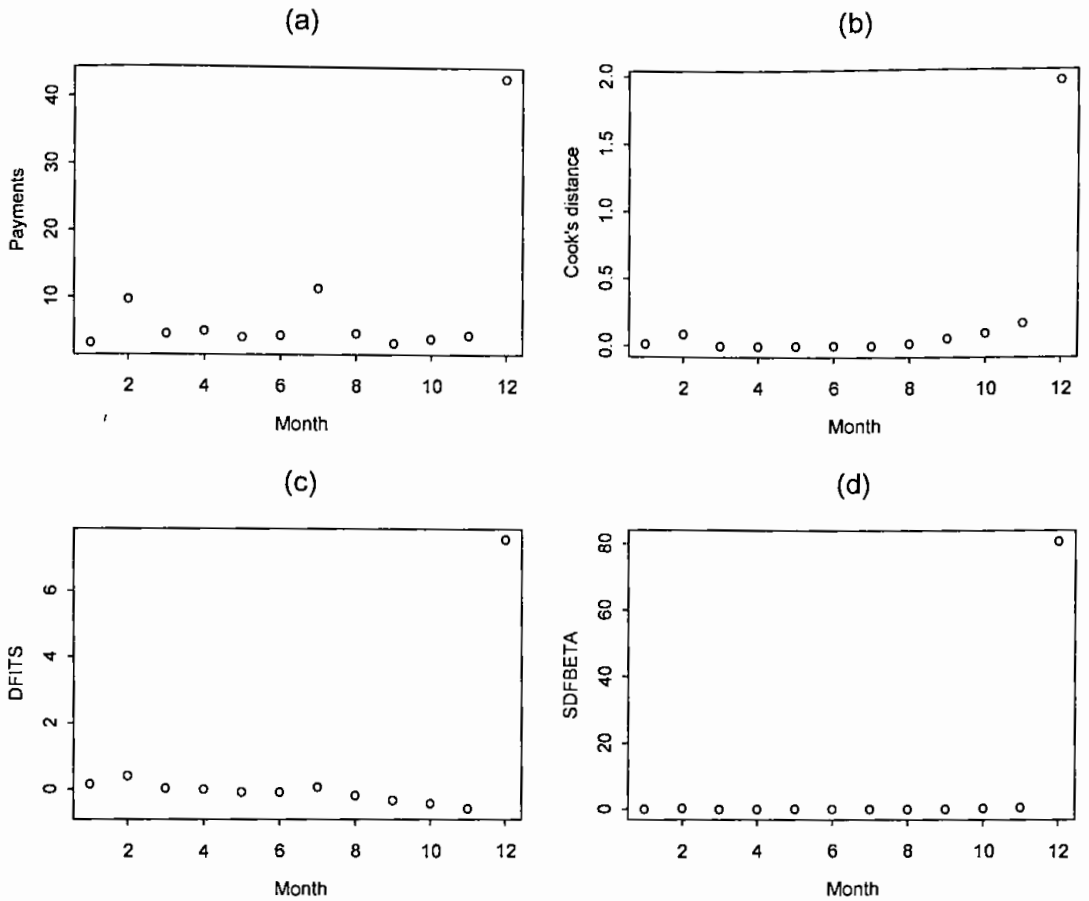


Figure 4.1 (a) Scatter plot Months versus Employment Payments; (b) Index plot of Cook's distance; (c) Index plot of DFFITS; (d) Index plot of SDFBETA

Hawkins et al. (1984) Data

To demonstrate the performance and comparison of the new measure with existing single case deletion diagnostic measures like Cook's distance and *DFFITS*, we analyze the Hawkins *et al.* (Hawkins *et al.* (1984)) artificial data as the identification of multiple influential observations. The data set consists of 75 observations in 4 dimensions, one for response variable others are explanatory variables. Rousseeuw and Leroy (1987) showed that observations 1-10 are ten high leverage outliers and the observations 11-14 are four (good leverage) points that are well accommodated by the LMS fit. Table 4.2 shows single case diagnostic results, Cooks distance identifies only observation 14 as influential

and *DFFITS* identifies 4 observations 11-14 as influential. At this stage we apply the newly proposed measure *SDFBETA* to identify the observations. According to our measure values consisting in column 6 of table 4.2, show observations 11-14 (four cases) are influential observations, which are as same as *DFFITS*. We know from the literature that first 10 observations are the most influential as they are at the same time, outliers and high leverage points, but they are masked by the next 4 observations (11-14). All three (*CD, DFFITS and SDFBETA*) are totally failed to identify multiple influential cases. Imon (2005) showed all first 14 observations as influential observations. We reach in conclusion that our single case deletion diagnostic measure can not identify all the influential cases properly as Cook's distance and *DFFITS*, because of masking effect. We want to improve our measure attach with the group deletion ideas like Hadi and Simonoff (1993), and Atkinson (1994) and generalize it for multiple linear regression diagnostic purpose.

Table 4.2 Single case diagnostics for first 14 cases of Hawkins *et al.*, (1984) data

Index	h_{ii} (0.120)	Std.res (2.500)	CD_i (1.000)	$\|DFFITS_i\|$ (0.693)	$\|SDFBETA_i\|$ (0.545)
1	0.063	1.552	0.040	0.406	0.176
2	0.060	1.831	0.053	0.470	0.235
3	0.086	1.396	0.046	0.430	0.202
4	0.081	1.187	0.031	0.352	0.135
5	0.073	1.413	0.039	0.399	0.172
6	0.076	1.588	0.052	0.459	0.228
7	0.068	2.077	0.079	0.575	0.354
8	0.063	1.762	0.052	0.464	0.230
9	0.080	1.255	0.034	0.372	0.150
10	0.087	1.413	0.048	0.439	0.211
11	0.094	**-3.657**	0.348	**-1.300**	**1.865**
12	**0.144**	**-4.501**	0.851	**-2.168**	**5.488**
13	**0.109**	**-2.881**	0.254	**-1.065**	**1.274**
14	**0.564**	**-2.558**	**2.114**	**-3.030**	**21.044**

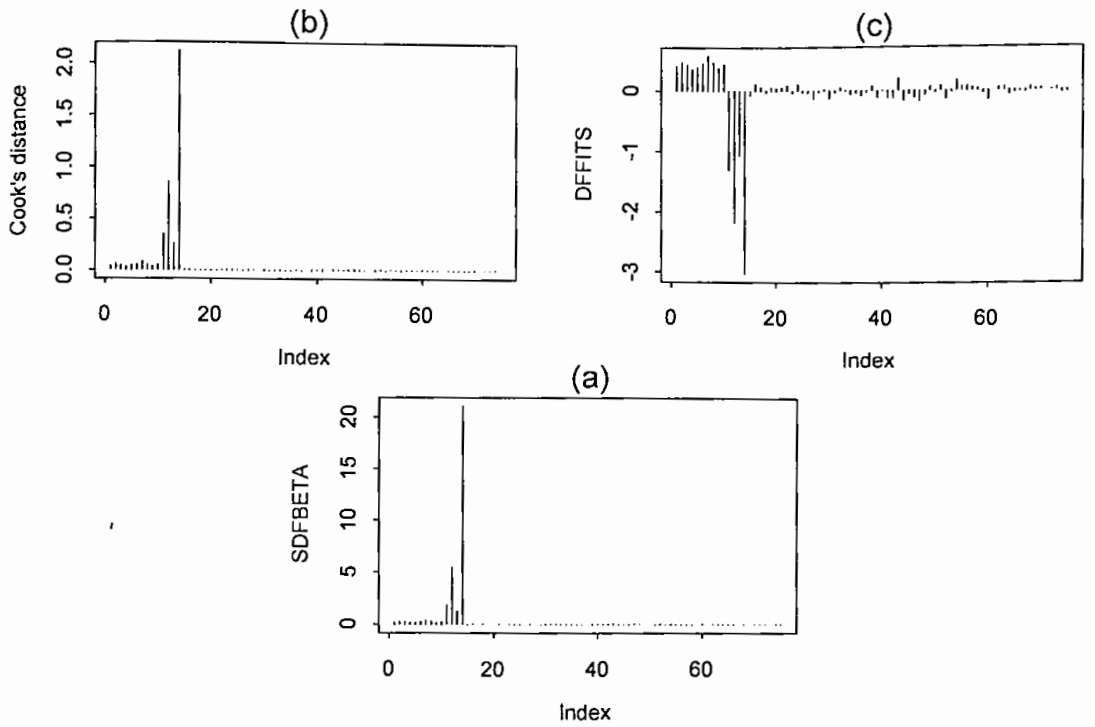


Figure 4.2 Index plot of (a) SDFBETA; (b) Cook's distance; (c) DFFITS

4.2 Generalized Squared Difference in Beta (*GSDFBETA*)

The diagnostic tool which is discussed in previous section designed for the identification of a single influential observation. But reality, hardly ever data sets contain just a single influential observation; a group of influential observations is present most of the times. It is now well known that a group of influential observations may distort the fitting of a model in such a way that influential observations have artificially very small residuals so that they may appear as inliers. Atkinson (1986) pointed out, "in the presence of masking single deletion diagnostic method fail to reveal outliers and influential observations." Therefore we need detection techniques for the identification of multiple influential observations and are free from masking and swamping phenomena.

Basic Philosophy

A good number of diagnostic methods are available in literature and some of them are placed in the previous sections. Most of them for the identification of multiple influential observations, attempt to separate the data into a 'clean' subset without any types of outliers and a complementary subset of the observations that contains all the potential outliers. It is generally used robust techniques and/or popular diagnostic techniques like Cook's distance, *DFFITS*, *etc* for identifying suspects. To perform the task, first the group of potential outliers/unusual observations are deleted at a time and then measure sensitivity of the each of the observations by observing the difference between the estimates without the group of unusual observations and the estimates adding another observation one after another from the remaining set of observations.

4.2.1 *GSDFBETA* Algorithm

In this section, we introduce a group-deleted version of diagnostic measure, generalization of squared difference in beta (*SDFBETA*), to develop effective diagnostic tool for the identification of multiple influential observations in linear regression. We name this *GSDFBETA*. We suggest here to use robust techniques for the identification of suspects to get the advantages of robustness. We assume that d observations among a set of n observations are suspects and are deleted from the data set. Let us denote the set of cases 'remaining' in the analysis by R and the set of cases 'deleted' by D. Hence R contains $(n\text{-}d)$ cases (each case is closely associated with a single row of the data matrix X and the corresponding element of Y) after d cases in D are deleted. Without loss of generality, we assume that these observations are the last of d rows of X and Y and V is a variance-covariance matrix, so that we may rearrange data matrix as

$$X=\begin{bmatrix} X_R \\ X_D \end{bmatrix} \qquad Y=\begin{bmatrix} Y_R \\ Y_D \end{bmatrix} \qquad V=\begin{bmatrix} V_R & 0 \\ 0 & V_D \end{bmatrix}.$$

Let $\hat{\beta}_{(R)}$ be the corresponding vector of estimated coefficients when a group of observations indexed by D is omitted. Now by using the idea of Hadi and Simonoff (1993) we define our proposed measure *GSDFBETA* as

$$GSDFBETA_i = \begin{cases} \dfrac{\left(\hat{\beta}_{(R)} - \hat{\beta}_{(R-i)}\right)^T X_R^T X_R \left(\hat{\beta}_{(R)} - \hat{\beta}_{(R-i)}\right)}{\hat{\sigma}^2_{(R-i)}\left(1 - h_{ii(R)}\right)}, & i \in R \\ \dfrac{\left(\hat{\beta}_{(R+i)} - \hat{\beta}_{(R)}\right)^T X_D^T X_D \left(\hat{\beta}_{(R+i)} - \hat{\beta}_{(R)}\right)}{\hat{\sigma}^2_R\left(1 - h_{ii(R+i)}\right)}, & i \notin R \end{cases} \tag{4.13}$$

where

$$\hat{\beta}_{(R)} = \left(X_R^T X_R\right)^{-1} X_R^T Y_R, \tag{4.14}$$

$$h_{ii(R+i)} = x_i^T \left(X_R^T X_R + x_i x_i^T\right)^{-1} x_i = \frac{h_{ii(R)}}{1 + h_{ii(R)}}, \tag{4.15}$$

$$\begin{aligned} \hat{\beta}_{(R-i)} &= \left(X_R{}^T X_R - x_i x_i{}^T\right)^{-1}\left(X_R{}^T Y_R - x_i y_i\right) \\ &= \hat{\beta}_{(R)} - \frac{\left(X_R{}^T X_R\right)^{-1} x_i}{1 - h_{ii(R)}} \hat{\varepsilon}_{i(R)}, \end{aligned} \tag{4.16}$$

$$\begin{aligned} \hat{\beta}_{(R+i)} &= \left(X_R{}^T X_R + x_i x_i{}^T\right)^{-1}\left(X_R{}^T Y_R + x_i y_i\right) \\ &= \hat{\beta}_{(R)} - \frac{\left(X_R{}^T X_R\right)^{-1} x_i}{1 + h_{ii(R)}} \hat{\varepsilon}_{i(R)}, \end{aligned} \tag{4.17}$$

where

$$\hat{\varepsilon}_{i(R)} = y_i - x_i^T \hat{\beta}_{(R)}.$$

4.2.2 Relation with *GDFFITS* and *GSDFBETA*

Imon (2005) developed generalized *DFFITS* (*GDFFITS*) for the identification of multiple influential observations in linear regression as

$$GDFFITS_i = \begin{cases} \dfrac{\hat{y}_{i(R)} - \hat{y}_{i(R-i)}}{\hat{\sigma}_{(R-i)}\sqrt{h_{ii(R)}}}, & i \in R \\ \dfrac{\hat{y}_{i(R+i)} - \hat{y}_{i(R)}}{\hat{\sigma}_{(R)}\sqrt{h_{ii(R+i)}}}, & i \notin R \end{cases} \tag{4.18}$$

where $$\hat{y}_{i(R+i)} = x_i^T \hat{\beta}_{(R+i)} = \hat{y}_{i(R)} + \frac{h_{ii(R)}}{1+h_{ii(R)}} \hat{\varepsilon}_{i(R)} \tag{4.19}$$

$$h_{ii(R)} = x_i^T (X_R^T X_R)^{-1} x_i \ ; i = 1, 2, ..., n$$

$$h_{ii(R+i)} = \frac{h_{ii(R)}}{1+h_{ii(R)}}.$$

Imon (2005) suggests considering observations as influential if

$$\left|GDFFITS_i\right| \geq 3\sqrt{k/(n-d)}\,. \tag{4.20}$$

The above results of *GDFFITS*, together with results of (4.15), (4.17) and (4.19), helps us to re-express *GDFFITS* in terms of group deletion residuals and leverages as

$$GDFFITS_i = \begin{cases} \sqrt{\dfrac{h_{ii(R)}}{1-h_{ii(R)}}} \dfrac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1-h_{ii(R)}}} & for \quad i \in R \\ \sqrt{\dfrac{h_{ii(R)}}{1+h_{ii(R)}}} \dfrac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_{R}\sqrt{1+h_{ii(R)}}} & for \quad i \notin R \end{cases}. \tag{4.21}$$

We may call the leverage components of *GDFFITS* as shown in equation (4.21) as generalized weights. These weights are denoted by h_{ii}^* and defined as

$$h_{ii}^* = \begin{cases} \dfrac{h_{ii(R)}}{1-h_{ii(R)}} & for \quad i \in R \\ = \dfrac{h_{ii(R)}}{1+h_{ii(R)}} & for \quad i \notin R \end{cases}. \tag{4.22}$$

Thus, *GDFFITS* values can also be re-expressed in terms of generalized Studentized residuals and generalized weights as

$$GDFFITS_i = \sqrt{h_{ii}^*}\, t_i^*, \qquad i = 1,2,\cdots,n \tag{4.23}$$

where t_i^* are the generalized Studentized residuals and defined as

$$t_i^* = \begin{cases} \dfrac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1-h_{ii(R)}}} & for \quad i \in R \\ \dfrac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_{R}\sqrt{1+h_{ii(R)}}} & for \quad i \notin R \end{cases}. \tag{4.24}$$

Now *GDFFITS* can be re-expressed as

$$GDFFITS_i^2 = \begin{cases} \dfrac{\left(\hat{\beta}_{(R)} - \hat{\beta}_{(R-i)}\right)^T X_R{}^T X_R \left(\hat{\beta}_{(R)} - \hat{\beta}_{(R-i)}\right)}{\hat{\sigma}^2_{(R-i)} h_{ii(R)}}, & i \in R \\ \dfrac{\left(\hat{\beta}_{(R+i)} - \hat{\beta}_{(R)}\right)^T X_D{}^T X_D \left(\hat{\beta}_{(R+i)} - \hat{\beta}_{(R)}\right)}{\hat{\sigma}^2_{(R)} h_{ii(R+i)}}, & i \notin R \end{cases} . \quad (4.25)$$

After some calculations we involve the equations (4.13) and (4.24) to make the relation as follows.

$$GSDFBETA_i = \begin{cases} \dfrac{(GDFFITS_i)^2}{1 - h_{ii(R)}} h_{ii(R)}, & i \in R \\ (GDFFITS_i)^2 h_{ii(R)}, & i \notin R \end{cases} . \quad (4.26)$$

Finally, we define *GSDFBETA* (by using equation (4.22) and (4.25)) for calculation purpose as

$$GSDFBETA_i = \begin{cases} \dfrac{h_{ii}{}^* \left(t_i^*\right)^2}{1 - h_{ii(R)}} h_{ii(R)} & i \in R \\ h_{ii}{}^* \left(t_i^*\right)^2 h_{ii(R)} & i \notin R \end{cases} . \quad (4.27)$$

4.2.3 Cutoff Value

Observations corresponding to large *GSDFBETA* are declared as influential observations. We consider it (*i*-th observation) large and take as influential if it exceeds the following cut-off value. With the help of cut-off points from Imon (2005) and Vellman and Welsch (1981), we treat cut-off value for *GSDFBETA* as

$$\begin{aligned} |GSDFBETA_i| &\geq (cutoff \quad GDFFITS)^2 h_{ii(R)} \\ &= \left(3\sqrt{k/(n-d)}\right)^2 \frac{3p}{n-d} = \frac{27kp}{(n-d)^2} . \end{aligned} \quad (4.28)$$

4.2.4 Examples

In this section we consider two well-known data sets, which are referred for the identification of multiple influential observations in linear regression. We show and compare the performance of proposed *GSDFBETA* with *GDFFITS* as the group-deletion identification technique of multiple influential observations.

Hawkins et al. (1984) Data

Now we apply the proposed algorithm to compute *GSDFBETA* for the Hawkins *et al.* (1984) data set. Imon (2005) showed generalized Cook's distance (GCD) is totally failed to identify influential observations for the masking effects. Table 4.3 shows our proposed method and *GDFFITS* both can identify 14 (1-14) influential observations successfully. But the figures 4.3 and 4.4 show better performance of *GSDFBETA* compare to *GDFFITS* in sense of smoothness in regular observations and makes significant distances between regular and influential observations. *GSDFBETA* shows better homogeneity among the regular observations and very sensitive to influential observations.

Table 4.3 Group deletion measures of influence for Hawkins *et al.* (1984) data

Ind.	|GDFFITS| (0.768)	GSDFBETA (0.087)	Ind.	|GDFFITS| (0.768)	GSDFBETA (0.087)	Ind.	|GDFFITS| (0.768)	GSDFBETA (0.087)
1	5.177	387.586	26	-0.298	0.006	51	0.259	0.004
2	5.272	423.015	27	-0.342	0.011	52	-0.334	0.011
3	5.169	453.267	28	0.137	0.001	53	0.733	0.075
4	4.759	408.008	29	0.111	0.000	54	0.340	0.010
5	5.003	435.023	30	-0.005	0.000	55	0.019	0.000
6	5.151	414.234	31	-0.045	0.000	56	0.026	0.000
7	5.475	470.745	32	-0.207	0.003	57	0.269	0.004
8	5.410	433.614	33	-0.216	0.002	58	-0.059	0.000
9	4.899	408.732	34	-0.363	0.012	59	-0.036	0.000
10	5.149	423.540	35	0.216	0.004	60	-0.334	0.011
11	0.926	19.198	36	-0.287	0.003	61	-0.021	0.000
12	0.884	18.769	37	-0.198	0.004	62	0.324	0.009
13	1.171	31.186	38	0.356	0.007	63	-0.187	0.003
14	0.857	20.670	39	-0.349	0.009	64	-0.203	0.003
15	-0.229	0.005	40	-0.145	0.001	65	0.341	0.007
16	0.255	0.007	41	-0.006	0.000	66	-0.357	0.007
17	-0.101	0.001	42	-0.219	0.004	67	-0.182	0.001
18	0.018	0.000	43	0.397	0.016	68	0.423	0.018
19	0.077	0.000	44	-0.262	0.006	69	0.055	0.000
20	0.237	0.005	45	-0.283	0.006	70	0.377	0.007
21	0.318	0.004	46	-0.098	0.001	71	0.097	0.000
22	0.222	0.004	47	-0.630	0.046	72	-0.029	0.000
23	-0.312	0.004	48	0.115	0.001	73	0.241	0.003
24	0.259	0.003	49	0.405	0.010	74	-0.326	0.006
25	-0.171	0.003	50	-0.105	0.001	75	0.267	0.007

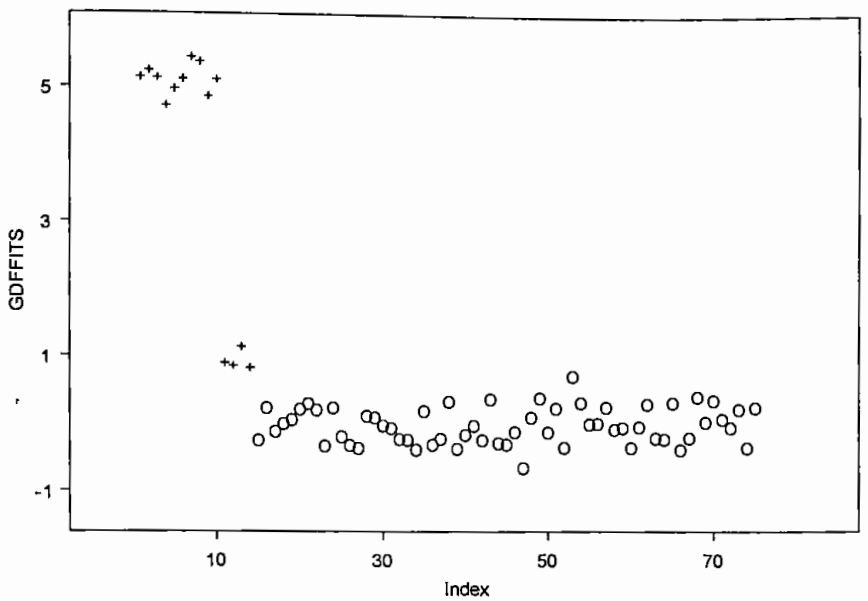


Figure 4.3 Index plot of GDFFITS for Hawkins *et al.* (1984) data

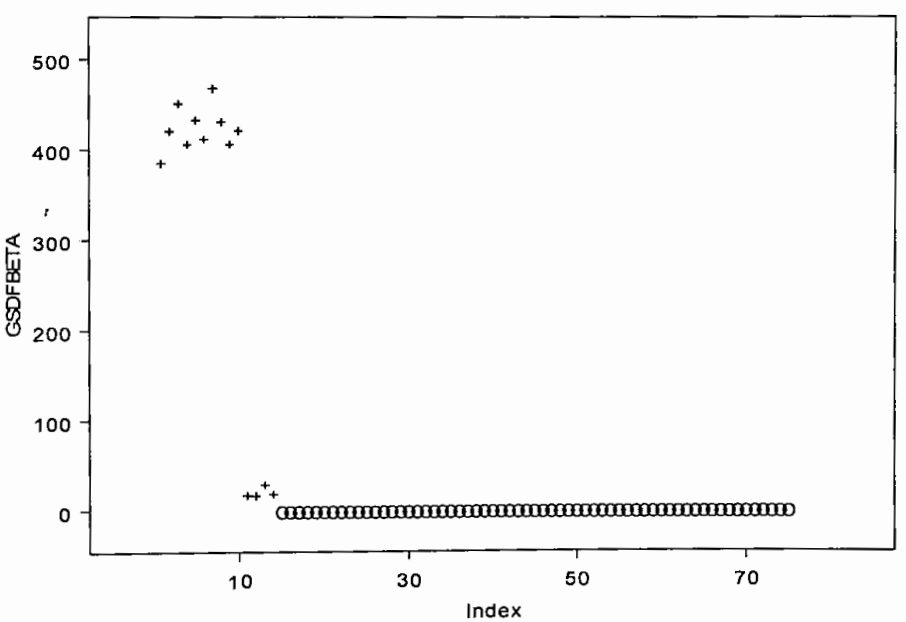


Figure 4.4 Index plot of GSDFBETA for Hawkins *et al.* (1984) data

Stack Loss Data

We consider the stack loss data presented by Brownlee (1965) that has been extensively analyzed in the diagnostic literature since then. This three-predictor (Air flow, Cooling water inlet temperature and Acid concentration and stack loss for response) data set contains 21 observations with five influential observations; three of them are high leverage outliers (cases 1, 3 and 21), one of them (case 4) is an outlier and another one (case 2) is a high leverage point. When the OLS technique is employed to the data we observe from table 4.4 that most of the traditional diagnostic methods fail to focus the influential cases. Cook's distance does not identify any one of the observations as influential. Studentized residuals (t_i) and *DFFITS* identify only one observation (case 21) as unusual as well as *SDFBETA*. Leverage values fail to identify genuine high leverage points but swamp in a good case (observation 17) as high leverage point. To compute *GSDFBETA*, we at first select suspect cases for deletion. The robust RLS technique identifies cases 1, 3, 4, and 21 as outliers. The rule based on generalized potential marks observations 1, 2, 3, and 21 as high leverage points. Thus our deletion set contains five different observations (cases 1, 2, 3, 4, and 21). When the regression model is fitted without these five points we observed from table 4.5 that all these observations have significant *GDFFITS* and *GSDFBETA* values that confirms our suspicion that these observations are influential. But *GCD* (Generalized Cook's Distance), another group deletion diagnostic measure, again fails to identify any of the influential observations.

We observe from figure 4.5 that *GDFFITS* correctly identifies all 14 influential observations. Figure 4.6 shows that in the index plot of *GSDFBETA* all influential observations are clearly identified and besides that all of them are far away from the usual ones. *GSDFBETA* shows better smoothness among the regular observations when we consider both the index plots at a glance.

Table 4.4 Single deletion measures for Stack loss data

Index	t_i(2.50)	w_{ii}(0.381)	CD(1.000)	|DFFITS|(0.873)	SDFBETA(1.067)
1	1.19	0.302	0.154	0.795	0.905
2	-0.72	0.318	0.060	-0.481	0.339
3	1.55	0.175	0.126	0.744	0.671
4	1.89	0.129	0.131	0.788	0.713
5	-0.54	0.052	0.004	-0.125	0.016
6	-0.97	0.077	0.020	-0.279	0.084
7	-0.83	0.219	0.049	-0.438	0.246
8	-0.48	0.219	0.017	-0.251	0.081
9	-1.05	0.140	0.045	-0.423	0.208
10	0.44	0.200	0.012	0.213	0.057
11	0.88	0.155	0.036	0.376	0.167
12	0.97	0.217	0.065	0.509	0.331
13	-0.48	0.158	0.011	-0.203	0.049
14	-0.02	0.206	0.000	-0.009	0.000
15	0.81	0.190	0.039	0.388	0.186
16	0.30	0.131	0.003	0.113	0.015
17	-0.61	**0.412**	0.065	-0.502	0.429
18	-0.15	0.161	0.001	-0.065	0.005
19	-0.20	0.175	0.002	-0.091	0.010
20	0.45	0.080	0.004	0.131	0.019
21	-2.64	0.285	0.692	**-2.100**	**6.168**

Table 4.5 Group deletion measures for Stack loss data

Index	GCD (1.00)	|GDFFITS| (1.500)	GSDFBETA (1.260)
1	0.154	**3.345**	**19.387**
2	0.060	**1.719**	**5.264**
3	0.126	**3.216**	**10.789**
4	0.131	**2.577**	**1.751**
5	0.004	0.142	0.004
6	0.020	-0.020	0.000
7	0.049	0.220	0.019
8	0.017	0.718	0.204
9	0.045	-1.099	0.281
10	0.012	0.526	0.106
11	0.036	0.407	0.041
12	0.065	0.315	0.043
13	0.011	-0.789	0.179
14	0.000	-0.575	0.093
15	0.039	0.113	0.005
16	0.003	-0.302	0.024
17	0.065	-0.308	0.100
18	0.001	-0.149	0.009
19	0.002	0.093	0.003
20	0.004	0.558	0.035
21	0.692	**-2.228**	**4.217**

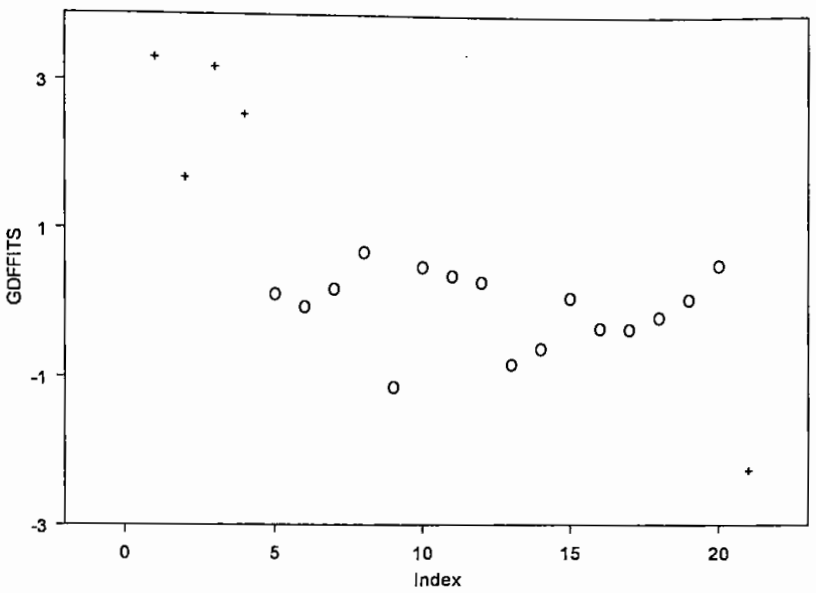


Figure 4.5 Index plot of GDFFITS for Stack loss data

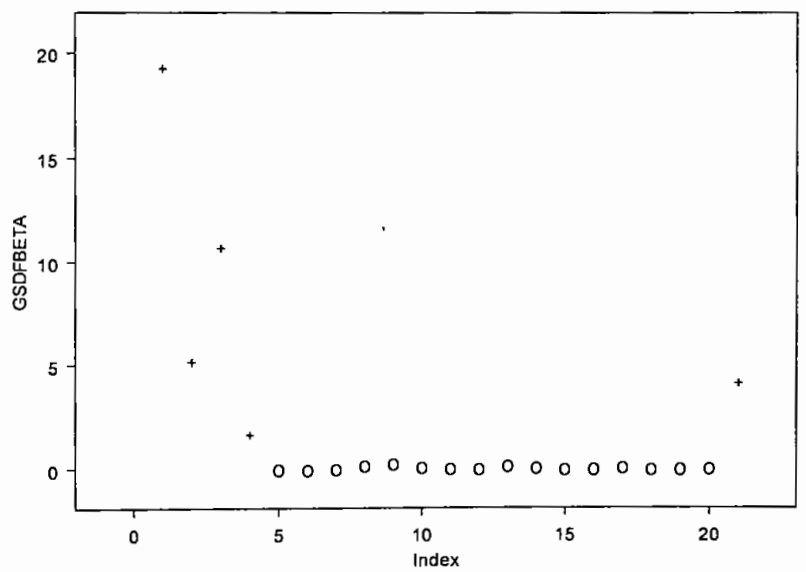


Figure 4.6 Index plot of GSDFBETA for Stack loss data

4.3 A New Measure for Identification of Multiple Influential Observations

It is a common practice in linear regression of measuring influence of an observation is to delete the case from the analysis and to investigate the norm of the change in the parameters or in the vector of forecasts resulting from this deletion. Pena (2005) introduced a new idea to measure the influence of an observation based on how this observation is being influenced by the rest of the data. In this section we would like to extend this idea to the group deletion technique suggested by Hadi and Simonoff (1993) and propose a new measure to identify influential observations in multiple linear regression. We investigate the usefulness of the proposed technique by two well-referred data sets and an artificial data with high-dimension, heterogeneous variances and large number of observations.

4.3.1 Pena (2005)'s Statistic

Pena (2005) introduced a new statistic totally in a different way, he mentioned, 'deletion of each sample point affects the forecast of a specific observation'. In his paper he outlined a procedure to measures how each sample point is being influenced by the rest of the data. He considered the vector

$$s_i = (\hat{y}_i - \hat{y}_i^{(-1)}, \ldots, \hat{y}_i - \hat{y}_i^{(-n)})^T, \tag{4.29}$$

where $\hat{y}_i^{(-i)}$ is the i-th fitted value when the i-th observation is deleted, and defined the statistic for the i-th observation as

$$S_i = \frac{s_i^T s_i}{p\, V(\hat{y}_i)}, \tag{4.30}$$

where p is the number of explanatory variables and $V(\hat{y}_i)$ is the variance of the i-th fitted value. This statistic can be re-expressed as

$$S_i = \frac{1}{p\hat{\sigma}^2 h_{ii}} \sum_{j=1}^{n} \frac{h_{ji}^2 \hat{\varepsilon}_j^2}{(1-h_{jj})^2}, \tag{4.31}$$

where h_{ji} is the ji-th element of the leverage matrix H,

$$\hat{y}_i - \hat{y}_i^{(-j)} = \frac{h_{ji}\hat{\varepsilon}_j}{1-h_{jj}}, \tag{4.32}$$

and

$$V(\hat{y}_i) = \hat{\sigma}^2 h_{ii}. \tag{4.33}$$

Pena proposed the observations with values of the statistic larger than

$$(S_i - E(S_i))/std(S_i) \tag{4.34}$$

as outliers and the observations having

$$S_i \geq median(S_i) + 4.5MAD(S_i)$$

or

$$S_i \leq \max(0, \quad median(S_i) - 4.5MAD(S_i)) \tag{4.35}$$

as heterogeneous observations. $E(S_i)$, $std(S_i)$ and $med(S_i)$ are the mean, standard deviation and median of the statistic respectively.

4.3.2 Our Proposed New Measure, M_i

From the beginning of the Cook's (1977) seminal article, most of the ideas of finding influential observations in regression are developed on the basis of 'deleting the observations one after another and measuring their effects on various aspects of the analysis.' It is now evident that the single case deletion techniques fail to detect multiple influential observations mainly because of masking and/or swamping problems. We develop an alternative way to look at how the forecast values are influenced by deleting each of the cases after deletion of suspected group of influential cases first. Basically we have accumulated the idea of Hadi and Simonoff (1993) with Pena (2005) and developed a measure M_i, which is expected to perform better when a group of influential observations exist in a data set. It is also effective to identify a

set of low leverage points. We show that the performance of the proposed technique is quite satisfactory in situations like clusters of high-leverage points and in large data sets in high dimensions with heterogeneous variances that are not easy to handle by the existing influence measures.

4.3.3 Algorithm of Proposed Measure

The proposed measure is a two-step measure based on Pena's (2005) idea attach with formal and/or informal group deletion methods. This helps us to reduce the maximum disturbance by deleting them at a time and "the deletion of which produces the largest reduction in the residual sum of squares" (Hadi and Simonoff, 1993). It helps to make the data more homogeneous than before. When we delete each sample point one after another, the forecasts of a specific observation would be more sensitive by the influential observations. Group deletion attaching with Pena's idea improves the measure and show nice results in presence of masking and/or swamping phenomena. Sometimes graphical display like index plot and/or character plot of explanatory and response variables could give us some prior ideas about the influential observations, but these plots are not useful for higher dimension of regressors. There are some suggestions in the literature (Atkinson, 1986; Rousseeuw and Leroy, 1987; and Rousseeuw and van Zomeren, 1990) for using robust regression techniques like least median of squares (LMS) (Rousseeuw, 1984), least trimmed squares regression (LTS) (Rousseeuw, 1985) and reweighted least squares (RLS), (Rousseeuw and Leroy, 1987) to find the suspect unusual/influential observations. Pena and Yohai (1995) introduced a method to identify influential subsets in linear regression by analyzing the eigen vectors corresponding to the non-null eigen values of an influence matrix. Here we consider one of the two proposed procedures of Hadi and Simonoff (1993), which was an adaptation of Hadi (1992, 1994), and is related to the proposals of Hawkins *et al.* (1984) and Atkinson and Mulira (1993). They suggest dividing the data set into two initial subsets: a 'basic' subset that contains the first $k+1$ clean observations and a 'nonbasic' subset that contains the remaining observations. In this purpose

some times they use an appropriate diagnostic measure and some times use clustering based backward-stepping method (Simonoff, 1991).

In our method, at the first step we find out all suspect unusual cases. We propose to use graphical display and/or regression diagnostic measure and/or robust regression methods whatever can helps to find the suspected group with maximum number of influential cases. In the second step, we assume that d observations among a set of n observations are suspected as influential observations and to be deleted. Let us denote a set of cases 'remaining' in the analysis by R and a set of cases 'deleted' by D. Hence without loss of generality, assume that these observations are the last d rows of X and Y so that

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix} \qquad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix}.$$

After formation of the deletion set indexed by D, we would compute the fitted values $\hat{y}^{(-D)}$ after d observations are deleted. Let $\hat{\beta}^{(-D)}$ be the corresponding vector of estimated coefficients when a group of observations indexed by D is omitted. We define the vector of difference between $\hat{y}_j^{(-D)}$ (j-th fitted value after the deletion of suspect group D) and $\hat{y}_{j(i)}^{(-D)}$ (j-th fitted value when the i-th observation is deleted after the deletion of D group first) as

$$t_{(i)}^{(-D)} = (\hat{y}_1^{(-D)} - \hat{y}_{1(i)}^{(-D)}, \ldots, \hat{y}_n^{(-D)} - \hat{y}_{n(i)}^{(-D)})^T \tag{4.36}$$

$$= (t_{1(i)}^{(-D)}, \ldots, t_{n(i)}^{(-D)})^T \tag{4.37}$$

and

$$t_{j(i)}^{(-D)} = \hat{y}_j^{(-D)} - \hat{y}_{j(i)}^{(-D)}$$

$$= \frac{h_{ji}\hat{\varepsilon}_i^{(-D)}}{1-h_{ii}}, j = 1,2,\ldots,n \tag{4.38}$$

where $h_{ji} = x_j^T (X^T X)^{-1} x_i$ and $\hat{\varepsilon}_i^{(-D)} = y_i - \hat{y}_i^{(-D)}$.

Finally, we introduce our new measure as squared standardized norm, name this M_i and define as

$$M_i = \frac{t_{(i)}^{(-D)^T} t_{(i)}^{(-D)}}{kV(\hat{y}_i^{(-D)})}, \tag{4.39}$$

where $$V(\hat{y}_i^{(-D)}) = s^2 h_{ii}; \quad and \quad s^2 = \frac{\hat{\varepsilon}^{(-D)^T} \hat{\varepsilon}^{(-D)}}{n-k}. \tag{4.40}$$

Using (4.36) to (4.39), we obtain

$$M_i = \frac{1}{ks^2 h_{ii}} \sum_{j=1}^{n} h_{ji}^2 \frac{\hat{\varepsilon}^2{}_i^{(-D)}}{(1-h_{ii})^2}. \tag{4.41}$$

It may not be easy to find a theoretical distribution of M_i. But in this situation it is a common practice (see Hadi, 1992) to use a confidence bound type cut-off value for it.

We consider i-th observation to be influential if it satisfies the rule

$$|M_i| > median(M_i) + 4.5\, MAD(M_i), \tag{4.42}$$

where $$MAD(M_i) = median\{|M_i - median(M_i)|\}/0.6745.$$

4.3.4 Examples

In this section we demonstrate the proposed measure and show the performance of our newly proposed measure in comparison with Pena's (2005) statistic and other two existing popular methods for the identification of influential observations in linear regression through several well-known data sets, which are frequently referred in the literature for studying the identification of influential observations, high leverage points and outliers.

Hertzsprung-Rusell Diagram Data

Our first example comes from Astronomy and the data is taken from the Hertzsprung-Rusell diagram of the star cluster CYG OBI, which contains 47 stars in the direction of Cygnus. Explanatory variable is the logarithm of the effective temperature at the surface of the star (Te), and Y is the logarithm of the light intensity (L/L_0). This data is given in Rousseeuw and Leroy (1987) and later analyzed by many authors as an interesting masking problem. The character plot of this data is shown in Figure 4.7 reveals that there are four outliers (cases 11, 20, 30, and 34) in the data and many authors consider other two points (7, 14) seem to be far from the regression line of the most of the data. Therefore, we consider all six observations as suspect unusual cases. We apply our proposed measure together with Cook's distance and *DFFITS* for this data and the results are presented in Table 4.6. We observe from this table Cook's distance fails to identify even a single influential observation, *DFFITS* can identify 4 cases (14, 20, 30, and 34), Pena's statistic identifies 10 observations (7, 11, 14, 17, 19, 20, 29, 30, 34, and 45) *i.e.*, this statistic is affected by swamping phenomena for the observations 17, 19, 29, and 45 . Our proposed measure (M_i) can detect 5 influential observations (7, 11, 20, 30, and 34) correctly. Figure 4.8 shows that the point 14 is close to LMS line than LS line, which supports that the case is not influential for analysis. Moreover, the M_i (proposed measure) as given in table 4.6 and figure 4.9 find out another influential observation 9, which was undetected before. The histogram of M_i (in figure 4.10) clearly indicates the presence of heterogeneity among the observations, which is not as much as clear in histogram of S_i (figure 4.10, (e)).

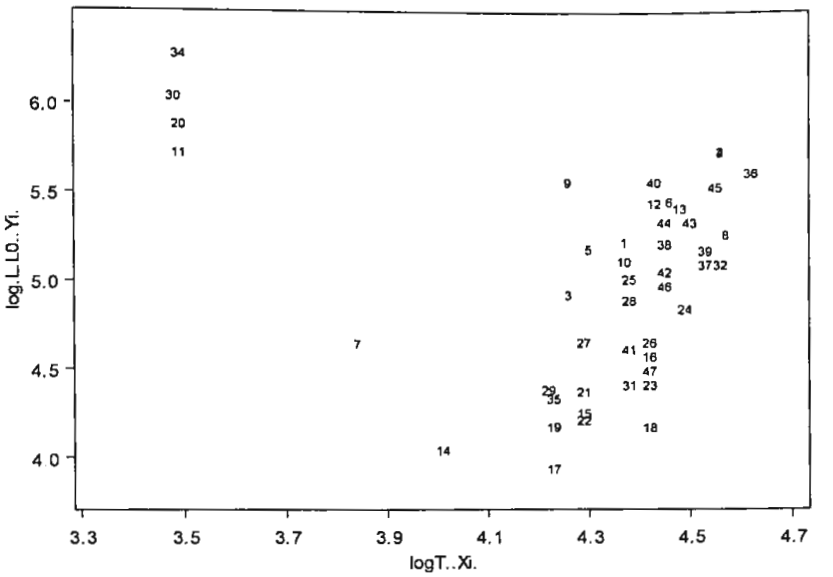


Figure 4.7 Character plot of the Hertzsprung-Rusell Diagram data

Table 4.6 Measures of influence for Hertzsprung-Rusell Diagram data

Ind.	CD (1.00)	|DFFITS| (0.412)	Si (0.254, 0.681)	Mi (-0.159, 0.234)	Ind.	CD (1.00)	|DFFITS| (0.412)	Si (0.254, 0.681)	Mi (-0.159, 0.234)
1	0.002	0.065	0.463	0.0569	25	0.000	0.010	0.455	0.0090
2	0.044	0.300	0.486	0.0431	26	0.004	-0.086	0.437	0.0375
3	0.000	-0.027	0.627	0.0653	27	0.005	-0.095	0.572	0.0016
4	0.044	0.300	0.486	0.0431	28	0.000	-0.022	0.455	0.0005
5	0.001	0.045	0.554	0.1154	29	0.017	-0.184	**0.705**	0.0000
6	0.012	0.152	0.437	0.0469	**30**	0.234	**0.691**	**1.044**	**8.0440**
7	0.045	-0.299	**1.039**	**0.7934**	31	0.012	-0.153	0.455	0.0706
8	0.009	0.131	0.493	0.0095	32	0.002	0.067	0.486	0.0334
9	0.010	0.144	0.627	**0.4091**	33	0.003	0.078	0.435	0.0080
10	0.001	0.035	0.463	0.0298	**34**	0.413	**0.935**	**1.044**	**8.8467**
11	0.067	0.365	**1.044**	**6.5866**	35	0.019	-0.195	0.686	0.0020
12	0.010	0.140	0.435	0.0697	36	0.043	0.296	0.528	0.0006
13	0.011	0.147	0.442	0.0245	37	0.002	0.060	0.467	0.0163
14	0.090	**-0.439**	**0.979**	0.0321	38	0.003	0.078	0.435	0.0080
15	0.020	-0.203	0.572	0.0401	39	0.004	0.086	0.467	0.0062
16	0.006	-0.109	0.437	0.0584	40	0.015	0.175	0.435	0.1130
17	0.046	-0.314	**0.686**	0.0814	41	0.005	-0.098	0.455	0.0212
18	0.025	-0.226	0.437	0.2327	42	0.000	0.031	0.435	0.0000
19	0.028	-0.241	**0.686**	0.0199	43	0.008	0.129	0.451	0.0052
20	0.136	**0.523**	**1.044**	**7.1984**	44	0.006	0.113	0.435	0.0262
21	0.014	-0.170	0.572	0.0165	45	0.024	0.220	0.479	0.0108
22	0.022	-0.214	0.572	0.0503	46	0.000	0.008	0.435	0.0030
23	0.012	-0.155	0.437	0.1143	47	0.009	-0.132	0.437	0.0840
24	0.000	-0.027	0.446	0.0425					

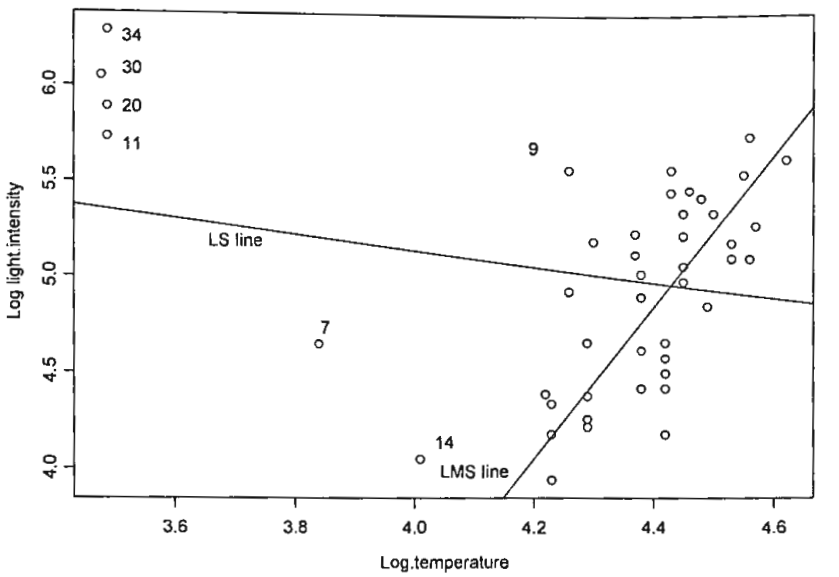
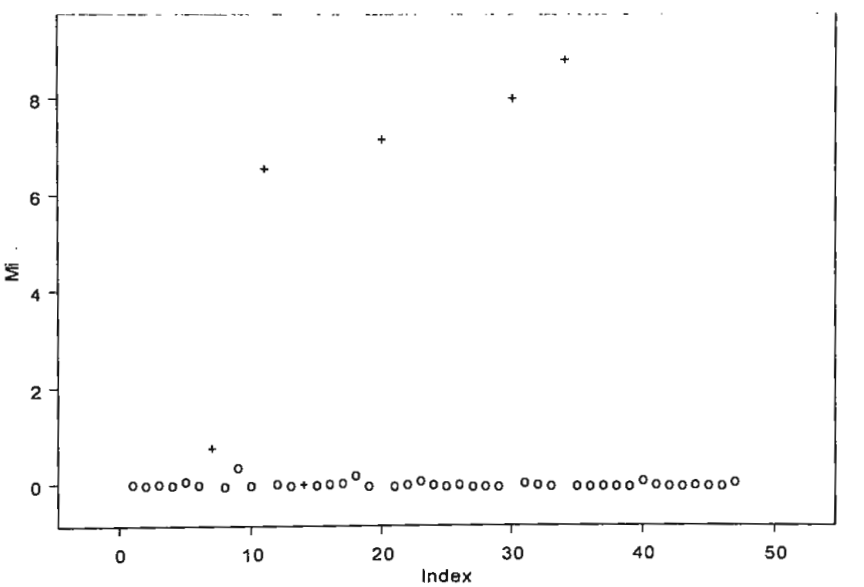


Figure 4.8 Log temperature Vs. log light intensity with LS and LMS line



Figurer 4.9 Index plot of proposed measure (M_i) for Hertzsprung-Rusell Diagram data

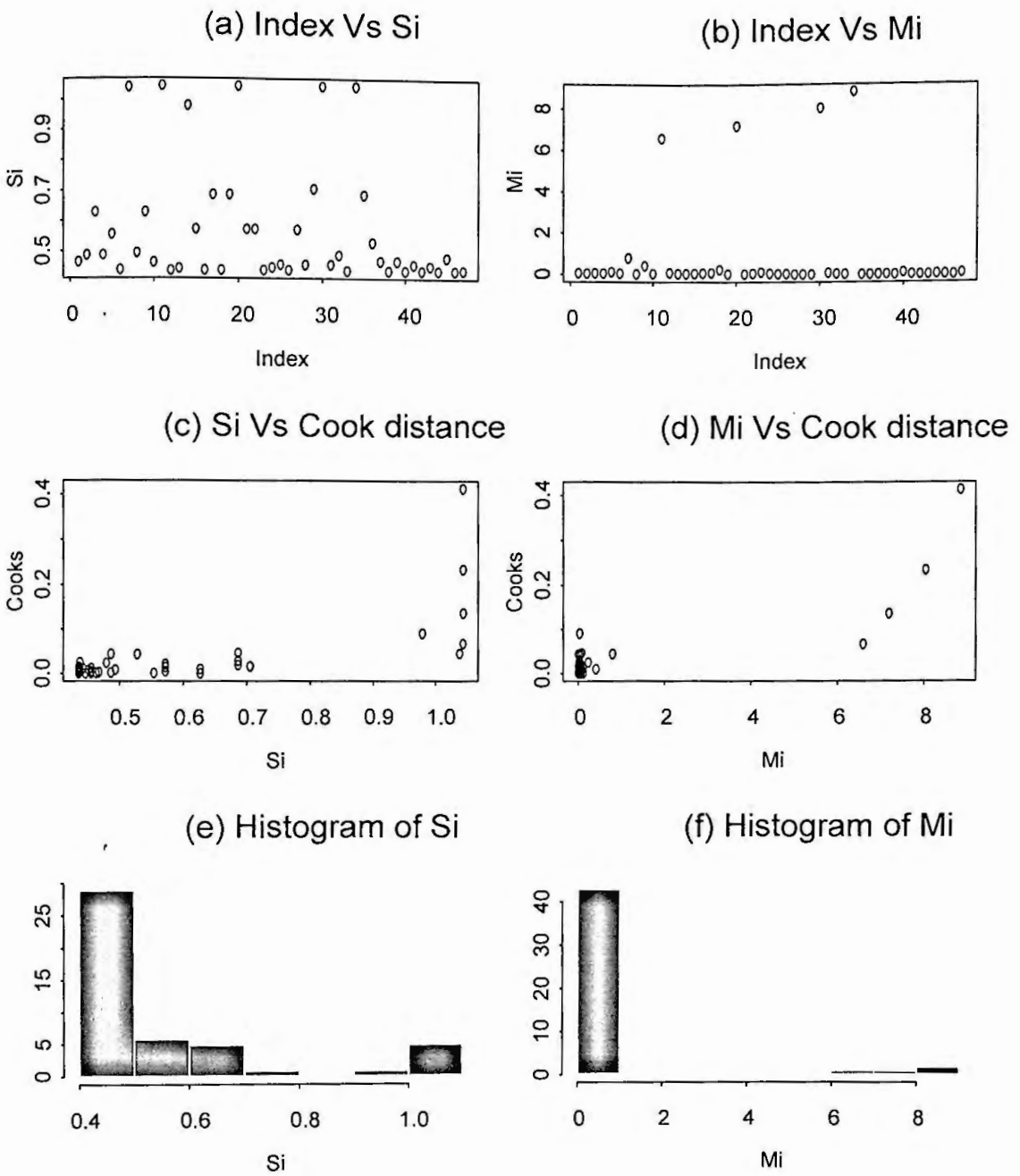


Figure 4.10 Index plots and Histograms of Influence measures for Hurtgsprung-Rusel Diagram data

Hawkins et al. (1984) Data

Hawkins *et al.* (1984) presents an artificial data set with three regressors containing 75 observations with 14 influential observations, among them 10 cases (1-10) are high leverage outliers and 4 cases (11-14) are high leverage points. Most of the single case deletion techniques fail to detect all of these influential observations properly. Some of them identify four high leverage points wrongly as outliers (Rousseeuw and Leroy, 1987). On the other hand, robust regression techniques like LMS and RLS identify outliers correctly, but they do not focus on the high leverage points and fails to identify them.

In our study we consider the first 14 data points as suspect influential cases and compute M_i measure based on the remaining 61. Table 4.7 shows these diagnostic measures together with Cook's distance, *DFFITS* and Pena (S_i). Here we observe that Cook's distance identifies only one (case 14) out of 14 influential observations. *DFFITS* identifies 7 observations (cases 2, 7, 8, 11, 12, 13 and 14) correctly but fails to detect other 7 cases. Figure 4.11 and table 4.7 indicate that Pena's measure identifies only one case (14) and greater variability is present among the values of the statistic. Table 4.7 shows our measure can successfully identify all the 14 influential cases. The merit of our method is supported also from Figure 4.12, where the index plots of M_i shows that all 14 influential cases are identified and are clearly separated from the other regular observations, in addition figure 4.12 shows comparatively less variability in the values of M_i than S_i. Histogram of M_i (Figure 4.13(d)) separates two groups of data clearly. The S_i versus Cook's distance and M_i versus Cook's distance give emphasis of a comparative performance between them, shows in figure 4.13 (e and f).

Table 4.7 Measures of influence for Hawkins *et al.* (1984) data

Ind.	CD (1.00)	|DFFITS| 0.462	Si (-1.47, 2.286)	Mi (-.019, 0.027)	Ind.	CD (1.00)	|DFFITS| 0.462	Si (-1.47, 2.286)	Mi(-.019, 0.027)
1	0.040	0.406	1.6004	**1.8060**	39	0.003	-0.109	0.7738	0.0058
2	0.053	**0.470**	1.8294	**1.9454**	40	0.000	-0.002	0.2092	0.0021
3	0.046	0.430	1.6034	**2.1760**	41	0.003	-0.118	0.3741	0.0000
4	0.031	0.352	1.6260	**1.9243**	42	0.004	-0.120	0.3463	0.0023
5	0.039	0.399	1.7113	**2.0261**	43	0.010	0.200	0.8728	0.0056
6	0.052	0.459	1.4338	**1.9643**	44	0.007	-0.168	1.3170	0.0026
7	0.079	**0.575**	1.6522	**2.1943**	45	0.001	-0.059	0.4717	0.0035
8	0.052	**0.464**	1.6748	**2.0148**	46	0.004	-0.127	1.2661	0.0004
9	0.034	0.372	1.5926	**1.9373**	47	0.008	-0.182	0.3769	0.0127
10	0.048	0.439	1.5000	**2.0530**	48	0.002	-0.081	0.4592	0.0006
11	0.348	**-1.300**	1.7053	**0.0914**	49	0.001	0.065	0.1148	0.0098
12	0.851	**-2.168**	1.6328	**0.0994**	50	0.000	-0.039	0.0683	0.0007
13	0.254	**-1.065**	1.9725	**0.1533**	51	0.002	0.077	0.1369	0.0042
14	**2.114**	**-3.030**	**2.3380**	**0.4167**	52	0.006	-0.148	0.2823	0.0043
15	0.001	-0.074	0.1486	0.0021	53	0.000	-0.016	0.6842	0.0144
16	0.003	0.114	0.2407	0.0024	54	0.006	0.159	1.1868	0.0050
17	0.001	0.059	0.3105	0.0004	55	0.001	0.061	0.5249	0.0000
18	0.000	-0.027	0.1740	0.0000	56	0.001	0.069	0.9839	0.0000
19	0.001	0.053	0.5737	0.0005	57	0.000	0.042	0.1922	0.0052
20	0.000	0.034	0.3163	0.0021	58	0.000	0.025	0.7025	0.0002
21	0.001	0.053	0.0830	0.0102	59	0.001	-0.045	0.1954	0.0001
22	0.002	0.093	0.5231	0.0025	60	0.006	-0.158	0.2783	0.0042
23	0.000	-0.033	0.4077	0.0087	61	0.000	-0.002	0.2602	0.0000
24	0.002	0.099	0.6894	0.0051	62	0.001	0.045	0.3167	0.0040
25	0.000	-0.020	0.4368	0.0011	63	0.001	0.056	1.0629	0.0016
26	0.000	-0.039	0.4395	0.0046	64	0.001	-0.065	0.1032	0.0018
27	0.004	-0.130	0.1132	0.0046	65	0.000	-0.023	1.0755	0.0070
28	0.000	-0.017	0.6837	0.0019	66	0.000	-0.023	0.2288	0.0085
29	0.000	0.024	0.1010	0.0011	67	0.000	-0.030	0.1875	0.0055
30	0.004	-0.127	1.0067	0.0000	68	0.001	0.054	0.3846	0.0064
31	0.000	-0.031	0.1057	0.0001	69	0.000	0.015	0.1470	0.0001
32	0.001	0.049	0.7879	0.0021	70	0.000	0.035	0.0888	0.0103
33	0.000	-0.008	0.8214	0.0037	71	0.000	0.001	0.0839	0.0010
34	0.001	-0.055	0.3766	0.0049	72	0.000	0.011	0.1358	0.0001
35	0.000	-0.034	0.1425	0.0020	73	0.000	0.042	0.0836	0.0045
36	0.002	-0.081	0.0664	0.0075	74	0.000	-0.040	0.1710	0.0069
37	0.000	-0.026	0.2053	0.0016	75	0.000	-0.041	0.5724	0.0028
38	0.002	0.082	0.1346	0.0080					

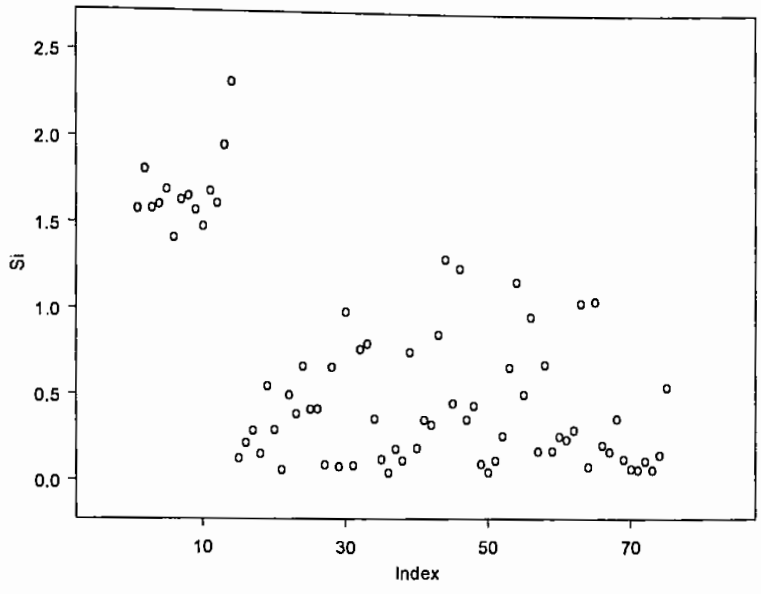


Figure 4.11 Index plot of S_i for Hawkins *et al.* (1984) data

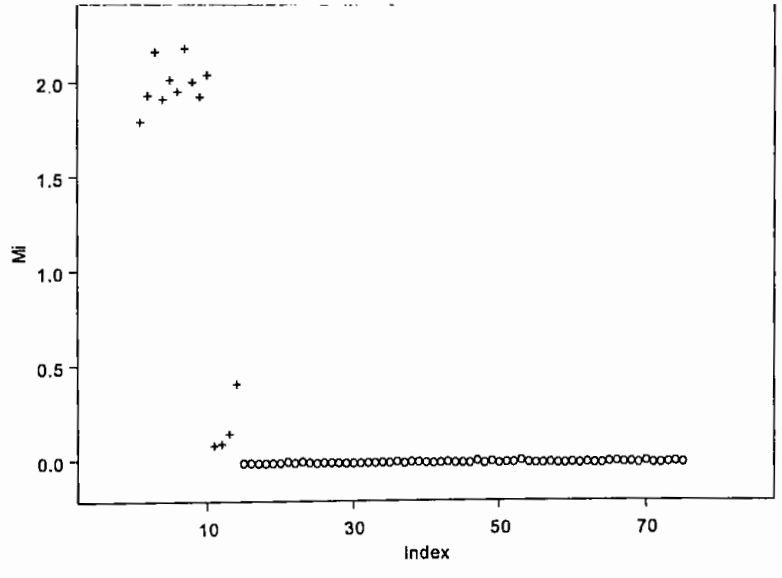


Figure 4.12 Index plot of M_i for Hawkins *et al.* (1984) data

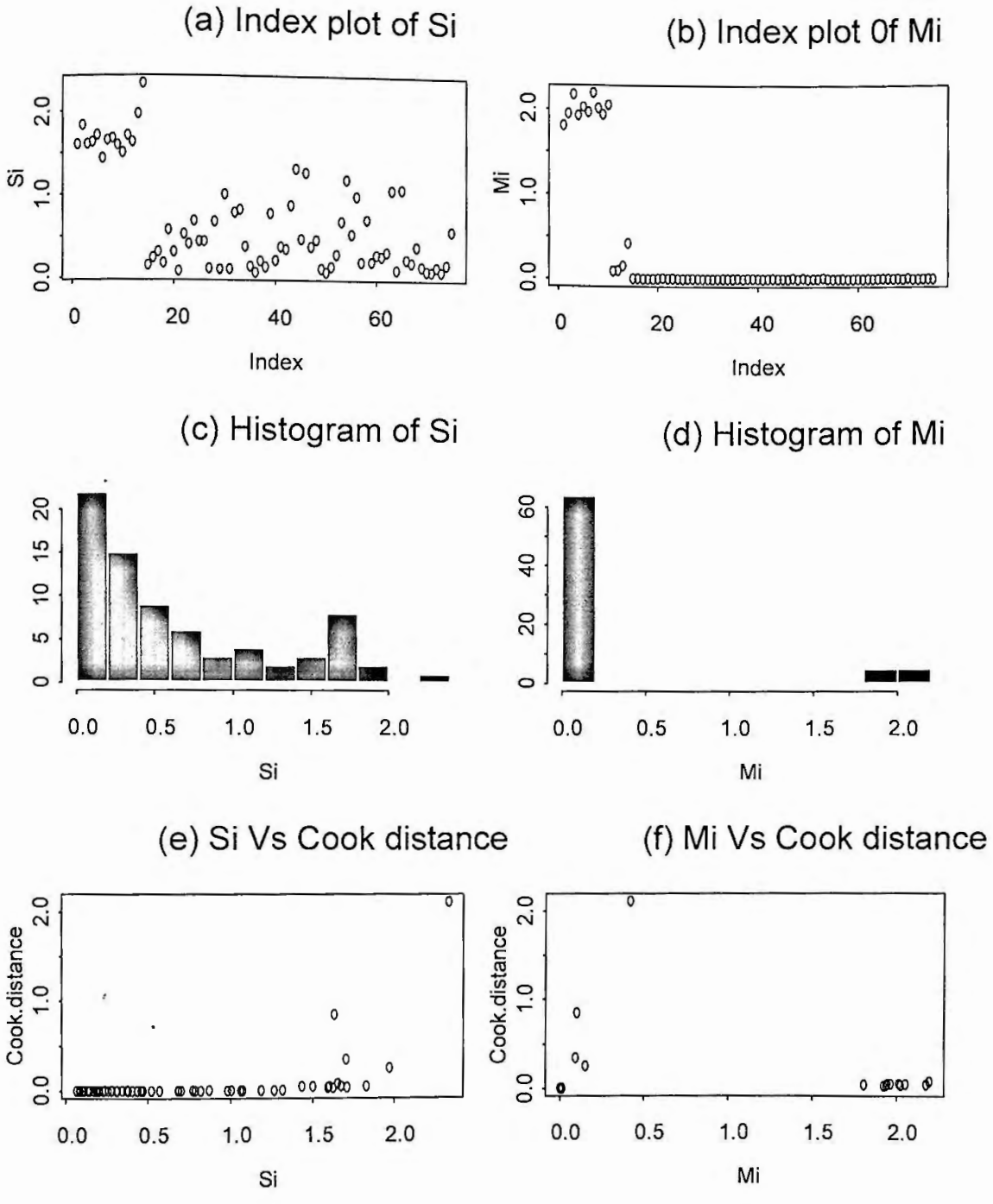


Figure 4.13 Influence measures for the Hawkins *et al.* (1984) data

High Dimensional, Large and Heterogeneous Artificial Data

Here we present an artificial data set that is generated in a similar fashion described by Pena (2005). The data set is a mixture of two regressors; heterogeneous and categorical variables generated by the following model.

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{20} X_{20} + \beta_{21} Z + \varepsilon \ . \qquad (4.43)$$

We generate 500 observations, where X's have 20 dimensions and they are independent random drawings from uniform distributions. The first 400 observations for each of the X variables are generated from Uniform (0, 10) and last 100 observations (401-500) from Uniform (9, 10), that makes the presence of heterogeneous variance in the data set. For the categorical variable Z, the first 400 observations are set at $Z = 0$ and the last 100 observations are set at $Z = 1$. For the null model we generate errors from Normal (0, 1). The parameter values have been chosen as $\beta_0 = \beta_1 = ... = \beta_{20} = 1$ and $\beta_{21} = -100$, so that the standard diagnostics of the regressing model does not show any evidence of heterogeneity.

Since the data set is large, results that we obtain from this artificial data are shown by the figure 4.14. The residual plots (scatter plot and histogram) show no indication of heterogeneity, but the residuals versus fitted values plot shows a clear indication of the presence of heterogeneity when the last 100 observations are deleted. We consider above mentioned 3 measures (Cook's distance, *DFFITS* and S_i) to identify the 100 influential cases but their index plots (figure 4.14 (e), (g) and figure 4.15(a)) show that they are totally failed to identify the influential cases. Figure 4.15 (a) and (b) shows Pena's measure identifies influential observations with some wrong indications. But our proposed M_i can successfully identify all influential cases (see figure 4.15 (c) and (d)). In figure 4.15(d), histogram of M_i also shows the heterogeneous variances in the data set that is not such clear in histogram of S_i (figure 4.15 (b)). The M_i versus cook's distance and M_i versus *DFFITS* (Figure 4.15 (f) and (h)) give the clear presentation of the effective performance of our proposed measure.

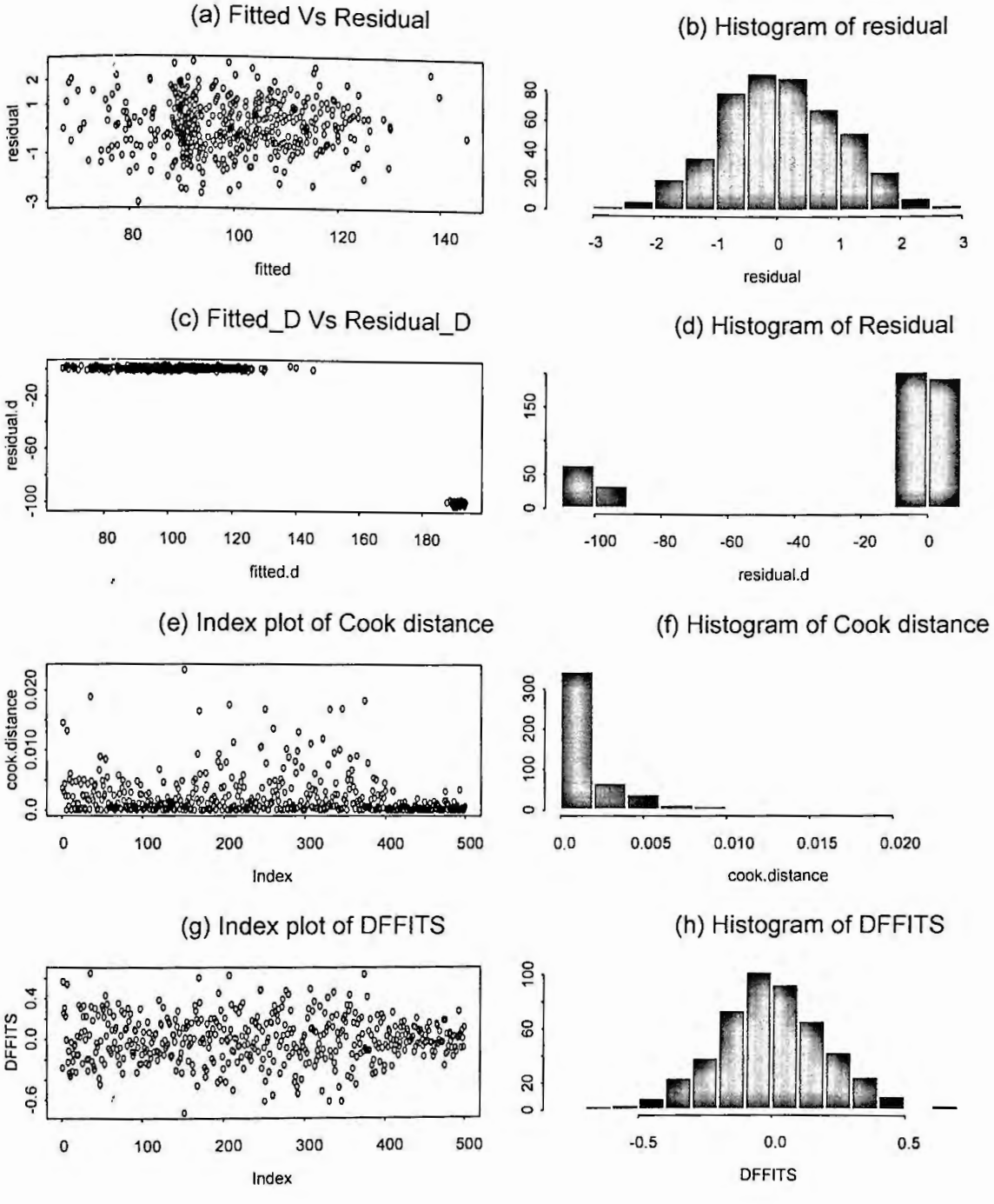


Figure 4.14 Influence measures plot for large-high dimensional artificial data

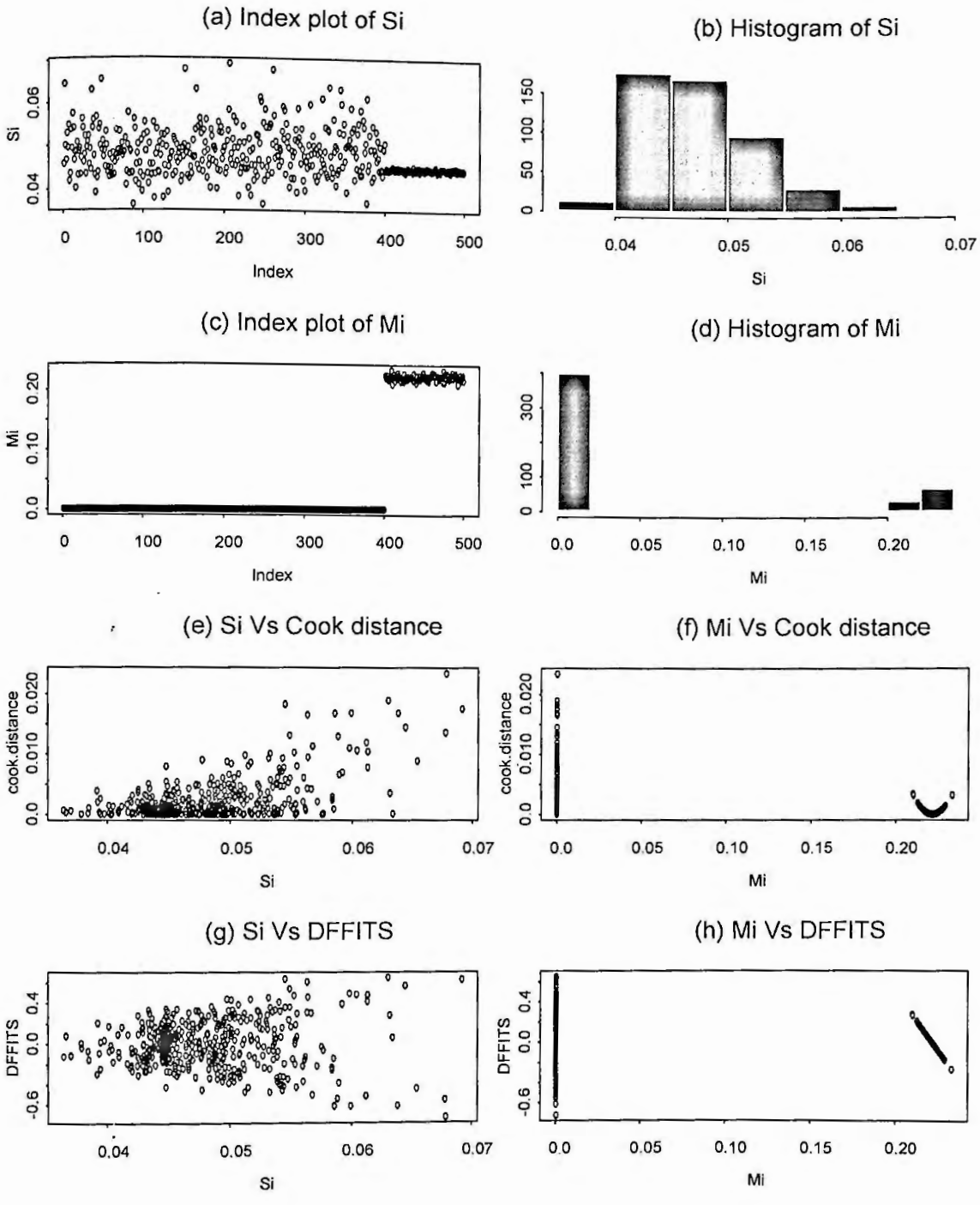


Figure 4.15 Influence measures plot for large-high dimensional artificial data

4.3.5 Simulation Results

In this section we show the performance of our measure (M_i) and compare with Pena's statistic (S_i) through simulation. We have done the simulation study for number of observations $n = 50$ and $n = 100$. For each of sample size we consider simple ($p = 1$) and multiple ($p = 5$) regression. The results are based on 1000 runs of each of the combination in which the contamination rates by unusuals are 40%, 30%, 20% and 10%.

Simulation of Simple Regression (p =1)

For simple linear regression we generate observations according to linear relation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $\beta_0 = 2, \quad \beta_1 = 1 \quad and \quad \varepsilon_i \sim N(0,.2)$.

For 'good' (G) observations $x_i \sim$ Uniform (1, 4) and a cluster of 'bad' (B) observations are generated which possesses a spherical bi-variate normal distribution with mean (7 and 2) and standard deviation 0.5.

Table 4.8

$n = 50, p = 1$, Runs = 1000
(CR = Correct Identification Rate,
IMC = Identification More than Contamination, *i.e.,* presence of swamping)

G/B(Contm.%)	Statistic	CR=100%	CR>75%	CR>50%	CR>25%	CR<25%	CR=0%	IMC
30/20	M_i	989	0	0	0	0	0	11
(40%)	S_i	601	99	20	14	8	258	0
35/15	M_i	822	0	0	0	0	0	178
(30%)	S_i	546	69	24	15	9	337	0
40/10	M_i	519	0	0	0	0	0	481
(20%)	S_i	105	3	5	3	3	881	0
45/05	M_i	192	0	0	0	0	0	808
(10%)	S_i	415	8	4	4	1	458	110

Table 4.9

$n = 100, p = 1$, Runs = 1000

(CR = Correct Identification Rate,

IMC = Identification More than Contamination, *i.e.*, presence of swamping)

G/B(Contm.%)	Statistic	CR=100%	CR>75%	CR>50%	CR>25%	CR<25%	CR=0%	IMC
60/40	M_i	977	0	0	0	0	0	23
(40%)	S_i	647	198	26	20	11	98	0
70/30	M_i	724	0	0	0	0	0	276
(30%)	S_i	587	110	40	29	30	204	0
80/20	M_i	294	0	0	0	0	0	706
(20%)	S_i	82	12	3	3	3	897	0
90/10	M_i	37	0	0	0	0	0	963
(10%)	S_i	419	10	7	3	3	499	59

Simulation of Multiple Regression (p=5)

We consider the multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i, \quad i = 1,2,...,n$$

where $\beta_0 = 2, \quad \beta_j = 1, \quad j = 1,2,...,5 \; and \quad \varepsilon_i \sim N(0,.2)$.

For 'good' (G) observations xij ~ Uniform (1, 4), and for 'bad' (B) observations

$$x_{ij} \in N(7,.5) \; and \quad y_i \in N(2,.5).$$

Table 4.10

$n = 50, p = 5$, Run = 1000

(CR = Correct Identification Rate,

IMC = Identification More than Contamination, *i.e.*, presence of swamping)

G/B(Con%	Statistic	CR=100%	CR>75%	CR>50%	CR>25%	CR<25%	CR=0%	IMC
30/20	M_i	923	0	0	0	0	0	77
(40%)	S_i	0	0	0	0	9	991	0
35/15	M_i	772	0	0	0	0	0	228
(30%)	S_i	0	0	1	0	7	992	0
40/10	M_i	596	0	0	0	0	0	404
(20%)	S_i	0	0	3	2	6	989	0
45/05	M_i	445	0	0	0	0	0	555
(10%)	S_i	0	2	3	8	24	962	1

Table 4.11

$n = 100, p = 5$, Run = 1000

(CR = Correct Identification Rate,

IMÇ = Identification More than Contamination, *i.e.*, presence of swamping)

G/B (Contm.%)	Statistic	CR=100%	CR>75%	CR>50%	CR>25%	CR<25%	CR=0%	IMC
60/40	M_i	901	0	0	0	0	0	99
(40%)	S_i	0	0	0	0	0	1000	0
70/30	M_i	735	0	0	0	0	0	265
(30%)	S_i	0	0	0	2	4	994	0
80/20	M_i	567	0	0	0	0	0	433
(20%)	S_i	0	5	11	13	21	950	0
90/10	M_i	427	0	0	0	0	0	573
(10%)	S_i	0	0	0	1	1	998	0

4.3.6 Findings from Simulation Study

Simulation tasks make some significant differences between the two: S_i and M_i as follows:

(a) M_i performs better than S_i in higher contamination rate.
(b) M_i performs better than S_i for multiple regression as well as simple regression.
(c) Most of the times S_i totally fails to identify unusual observations in multiple regressions.
(d) For lower contamination M_i affected by swamping.
(e) M_i does not matter sample size.

Chapter 5

"Humans are good, she knew, at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent."

Carl Sagon

Classification of Unusual Observations in Linear Regression

Present chapter proposes a five-fold plotting technique with a robust distance measure on a potential-residual (P-R) plot that can classify outliers, high leverage points and influential observations as well as identify them properly at the same time in a same graph. The proposed technique based on group deletion idea shows efficient performance in presence of masking and/or swamping phenomena. It shows its efficiency for multiple unusual observations by using three well-referred data sets and an artificial high-dimensional large data with heterogeneous variances.

5.1 Necessity of Classification

In linear regression, most of the time we try to diagnosis outliers and high leverage points that have very close ties with influential observations. It is a common belief that outliers would be highly influential, but that is not always true. Andrews and Pregibon (1978) showed that outlying observations might be little influence on the results. Their examples illustrated the existence of outliers that did not matter and high leverage points were likely to be influential. Chatterjee and Hadi (1986) showed that 'as with outliers, high leverage points need not be influential, and influential observations are not necessarily high leverage points'. Chatterjee and Hadi (1988) pointed out, "Patterns of the residuals are often more informative than their magnitudes. Graphical displays of residuals are,

therefore, much more useful than formal test procedures." We make the sense, graphical displays can help us to remove the problem of close ties of outliers and leverage values by visualizing the real patterns of residuals and leverage values. An extensive search (*e.g.*, Atkinson, (1981, 1985), Atkinson and Riani (2000), Barnett and Lewis (1995), Behnken and draper (1972), Cook and Weisberg (1982), Cook (1998), Rousseeuw and Leroy (1987), Rousseeuw and van Zomeren (1990), and Hubert *et al.* 2007) is still going for a reliable plotting procedure for identifying outliers and high-leverage points. Most of the diagnostic techniques in literature for identifying outliers and high-leverage points are focusing on both of them separately. Identification of both of them at the same time is necessary because presence of one makes the identification of the other very difficult (Pena and Yohai, 1995) and as a result proper identification of influential observations may not be possible. We propose a new type of potential-residual (P-R) plot that identifies outliers, high-leverage points and influential observations along with a Mahalanobis type robust distance measure and classify them properly at a time in a same five-fold graph, which is not possible by the plot analysis in literature and in existing statistical software packages.

5.2 Classification of Regression Data

In regression analysis observations (data) are classified basically into two: regular observations and unusual observations. Regular observations are those who have some specific pattern and good in number in the data set. On the contrary, observations are unusual in the sense that they are exceptional, they have extra role on model building process, or they may come from other population/s and do not follow the pattern of the majority of the data. Statisticians classify unusual observations into three: outliers, high-leverage points and influential observations. Following diagram makes the clear understanding about the classification in regression data.

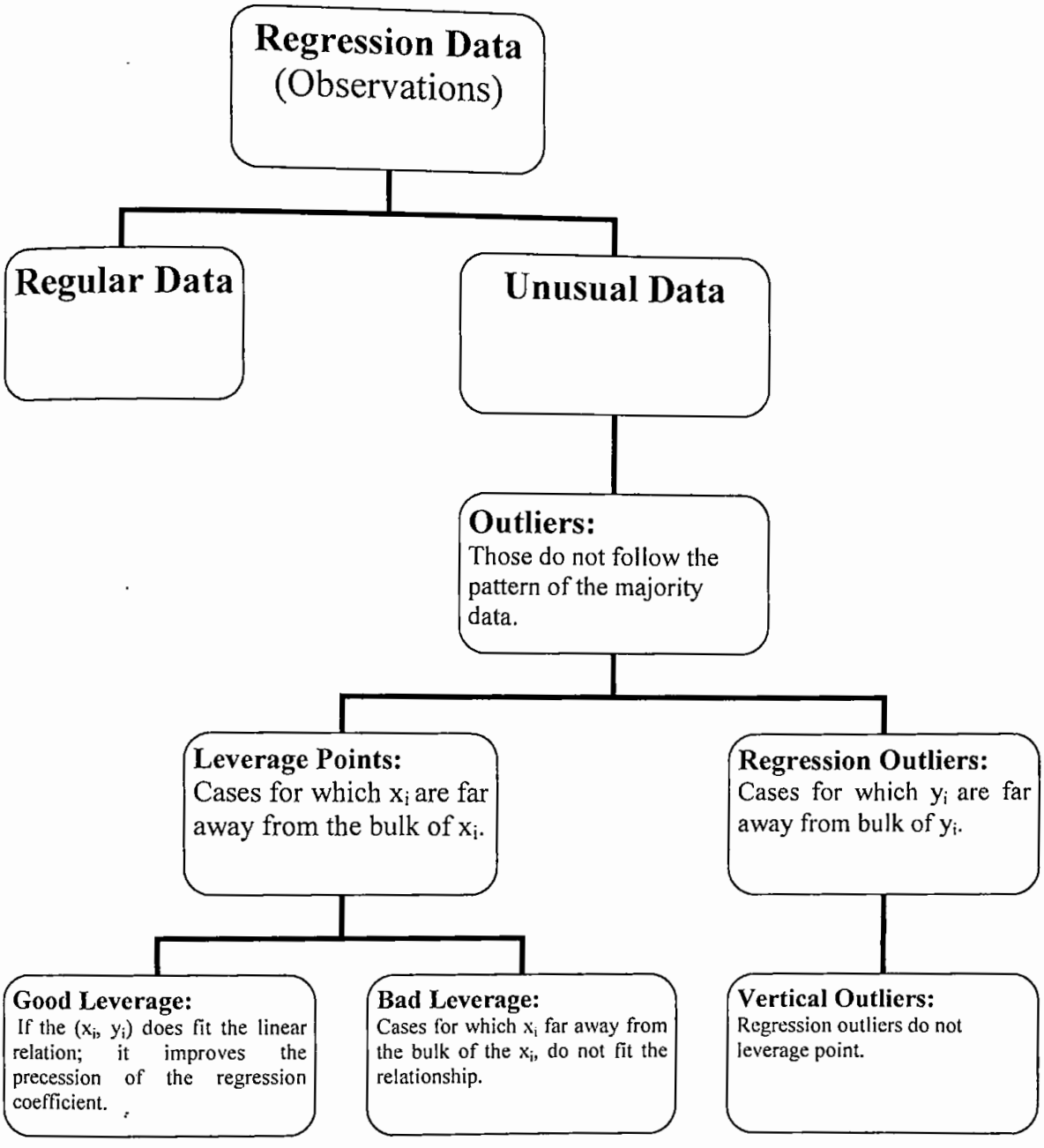


Figure 5.1 Classification of regression data

5.3 Classification through Identification

In regression diagnostics basic objective is to identify the influential observations that influence the model building process and decision making. Proper classification of unusual observations justifies identification of influential observations. In this regard we give a short discussion of identification techniques of outliers, high leverage points and influential observations successively.

5.3.1 Identification of Outliers

Outliers are generally identified by measuring the residual vector, r and is defined as

$$r = Y - X\hat{\beta} \tag{5.1}$$

$$= (I - H)\varepsilon, \tag{5.2}$$

In scalar form, i-th residual is

$$r_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j; \quad i = 1,2,...,n. \tag{5.3}$$

Clearly, if the h_{ij} are sufficiently small, r_i will serve as a reasonable alternative of ε_i.

Behnken and Draper (1972) use r_i for plots of residuals to identify outliers. Identity (5.2) indicates that the ordinary residuals are not independent (unless H is diagonal) and they do not have the same variance (unless the diagonal elements of H are equal). According to Montgomery *et al.* (2001), "non independence of the residuals has effect on their use for model adequacy checking as long as n is not small relative to the number of parameters, k". Scaled residuals are helpful for identifying outliers or extreme observations. Chatterjee and Hadi (1988) pointed out that the ordinary residuals are not appropriate for diagnostic purpose and a transformed version of them is preferable.

Since the approximate average variance of a residual is estimated by MS_{res} (mean squared residuals), a logical scaling for the residuals is the standardized residuals

$$r_i^* = \frac{r_i}{\sqrt{MS_{res}}}; i = 1,2,...,n \quad , \tag{5.4}$$

which have mean zero and approximate variance equal to one, consequently a large standardized residual potentially indicate an outlier. Since the residuals have different variances and are correlated, the variance of the i-th residual is

$$V(r_i) = \sigma^2(1-h_{ii}), \tag{5.5}$$

where σ is an estimate of MS_{res}. Daniel and Wood in 1971, introduced a type of (i-th internally Studentized) residual as

$$e_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}; i = 1,2,...,n, \tag{5.6}$$

when $\sigma_i = \hat{\sigma}\sqrt{1-h_{ii}}$. But many of the authors feel that internally Studentized residuals are over estimated by the extreme observations and they have suggested the i-th externally Studentized residuals,

$$e_i^* = \frac{r_i}{\hat{\sigma}^{(-i)}\sqrt{1-h_{ii}}}; \quad i = 1,2,...,n, \tag{5.7}$$

taking $\sigma_i = \hat{\sigma}^{(-i)}\sqrt{1-h_{ii}}$ where $\hat{\sigma}^{(-i)2} = \dfrac{Y^{(-i)^T}\left(I - H^{(-i)}\right)Y^{(-i)}}{n-k-1}$ $\quad i = 1,2,...,n,$

is the residual mean squared error estimate when the i-th observation is omitted and

$$H^{(-i)} = X^{(-i)}\left(X^{(-i)^T}X^{(-i)}\right)^{-1}X^{(-i)^T}, \quad i=1, 2,..., n,$$

is the prediction matrix for $X^{(-i)}$. Atkinson (1981) prefers e_i^* over e_i for detecting outliers. Imon (2005) proposed generalized Studentized residuals to coop up with multiple outliers by using group deletion and defined as

$$t_i^* = \begin{cases} \dfrac{r_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1-h_{ii(R)}}} & for \quad i \in R \\ \dfrac{r_{i(R)}}{\hat{\sigma}_R\sqrt{1+h_{ii(R)}}} & for \quad i \notin R \end{cases} \tag{5.8}$$

where $r_{i(R)} = y_i - x_i^T \hat{\beta}_{(R)}$, $\hat{\beta}_{(R)} = (X_R^T X_R)^{-1} X_R^T Y_R$, and $h_{ii(R)}$ is the i-th diagonal element of $H_{(R)}$ (where 'R' is the remaining group of observations after the deletion), which are analogous to residuals suggested by Hadi and Simonoff (1993) and Atkinson (1994).

5.3.2 Identification of High-Leverage Points

Observations corresponding to excessively large values of $h_{ii} = x_i^T (X^T X)^{-1} x_i$ are treated as high-leverage points. Among them twice-the-mean rule proposed by Hoaglin and Welsch (1978), thrice-the-mean rule proposed by Vellman and Welsch (1981) and suggestions of Huber (1981) are commonly used. Mahalanobis distance (1936) is the first distance measure used in multivariate analysis, also suggested to use as a measure of high-leverage points. Rousseeuw and van Zomeren (1990) showed the diagonal elements of hat matrix (H), h_{ii} has the positive relation with non-robust MD_i as

$$h_{ii} = \frac{1}{n-1} MD_i^2 + \frac{1}{n}. \tag{5.9}$$

Due to non robustness, MD sufferes from both masking and swamping. To remove the problems (masking, swamping) of multiple outliers Rousseeuw and van Zomeren (1990) proposed robust distance, by using minimum volume ellipsoid (*MVE,* Rousseeuw 1985). Hadi (1992) pointed out, "...in the presence of a high-leverage point the information matrix may breakdown and hence the observation may not have the appropriate leverage". In this connection, Hadi (1992) introduced potentials based on single case deletion and defined as

$$h_{ii(i)} = x_i^T (X_{(i)}{}^T X_{(i)})^{-1} x_i, \qquad i = 1,2,...,n, \tag{5.10}$$

where $X^{(-i)}$ is the matrix without the i-th observation. Observations corresponding to excessively large potential values are considered as high-leverage points. It is reported that the presence of multiple high leverage points may cause masking and/or swamping because of which some outliers and leverage points goes undetected (masking) and/or some innocent observations reveal as outliers or leverage points (swamping). Imon

(2002) suggested a group deletion measure (generalized potentials) for identifying multiple high-leverage points.

5.3.3 Identification of Influential Observations

Most popular identification techniques of influential observations, Cook's distance, DFFITS or DFBETAS consider both of outliers and leverage points together in a multiplicative form. Imon and Ali (2005) mentioned that, residuals together with leverage values may cause masking and swamping for which a good number of unusual observations remain undetected in the presence of multiple outliers and multiple high-leverage points. As a result identification of multiple influential observations would be troublesome. Hadi (1992) showed that the values of these statistics misleadingly small if either residuals or leverage values used in these statistics in a multiplicative form are small and consequently they could fail to identify potential outliers and leverage points. Hadi (1992) suggested a new type additive measure for identifying influential observations as

$$H_i^2 = \frac{k}{1-h_{ii}}\frac{d_i^2}{1-d_i^2} + \frac{h_{ii}}{1-h_{ii}}, \qquad i = 1,2,...,n, \tag{5.11}$$

where $d_i^2 = \frac{r_i^2}{r^T r}$ is the square of the i-th normalized residual and h_{ii} is the i-th diagonal elements of H for the full matrix. As a cut-off point, Hadi suggested H_i^2 to be large if it exceeds $median(H_i^2) + c\sqrt{mad(H_i^2)}$, c is appropriately chosen constant between 2 and 3. Imon and Ali (2005) mentioned that, since H_i^2 is a single case deletion diagnostic measure it may fail to identify multiple outliers and high-leverage points. They re-expressed the Hadi's (1992) statistic as

$$H_i^2 = \frac{k}{n-k-1}\frac{r_i^2}{\sigma^{(-i)^2}(1-h_{ii})} + \frac{h_{ii}}{1-h_{ii}} \qquad i = 1,2,...,n, \tag{5.12}$$

and finally they suggested another statistic as an additive form of residual and leverage and used group deletion idea (see Imon and Ali, 2005) as

$$ARL_i = k_1\left|t_i^*\right| + k_2 h_{ii}^* \qquad i = 1,2,\ldots,n, \tag{5.13}$$

where

$$t_i^* = \begin{cases} \dfrac{y_i - x_i^T \hat{\beta}_{(R)}}{\hat{\sigma}_R\sqrt{1-h_{ii(R)}}} & for \quad i \in R \\ \dfrac{y_i - x_i^T \hat{\beta}_R}{\hat{\sigma}_R\sqrt{1+h_{ii(R)}}} & for \quad i \notin R \end{cases} \tag{5.14}$$

$$h_{ii}^* = \begin{cases} \dfrac{h_{ii(R)}}{1-h_{ii(R)}} & for \quad i \in R \\ h_{ii(R)} & for \quad i \notin R \end{cases} \tag{5.15}$$

$k_1 = \sum_{i=1}^{n}\left|t_i^*\right|$ *and* $k_2 = \sum_{i=1}^{n} h_{ii}^*$. $\hat{\beta}_{(R)}$, $\hat{\sigma}_R$ *and* $h_{ii(R)}$ all carry the above and usual meaning.

They proposed to use confidence bound type cut-off point for ARL_i and gave as

$$ARL_i > median\ (ARL_i) + \ cMAD(ARL_i). \tag{5.16}$$

As usual c is any arbitrary chosen constant between 2 and 3. GDFFITS (Imon, 2005), a group deletion technique is also used for identifying multiple influential observations.

5.4 Proposed Method

We propose our technique by putting all together following remarkable statements from three leading regression diagnostic statisticians. Welsch (1982) pointed, "Neither the leverage nor the Studentized residual alone will usually be sufficient to identify influential cases". Hocking (1983) said, "I find that the diagonal elements of the hat matrix and 'deleted' Studentized residuals provide most of the evidence needed to track down maverick cases". Atkinson (1986) mentioned, "In presence of masking single deletion methods fail to reveal outliers and influential observations". We show the situations and effectiveness of potential versus residual (P-R) plot in a same space and try to keep away our measure from the affect of multiplicative phenomena. We define our technique in the following steps and the section '*Explanation*' clarifies the steps.

(i) Find the suspect (to be deleted) group of unusual observations.

(ii) Estimate the parameters after deleting the suspect group of cases.

(iii) By using the estimated parameters we calculate deleted residuals and we compute diagonal elements (group-deleted leverage/potentials) of the deleted leverage matrix.

(iv) We make a scatter plot (group-deleted potentials versus group-deleted standardized residuals (r_i)) and classify the observations according to pre assigned cut off points (which make boundary/confidence lines) that are given below.

(v) We draw a 95% joint confidence region of potentials and residuals for identifying influential observations in the same plot.

5.4.1 Explanation

A general approach of group deletion is to form a clean subset of the data that are presumably free of unusual observations and then test the outlyingness of the remaining points relative to the clean subset. Let M be the set of indexes of the observations in the clean sub set, and let Y_M and X_M be the subset of observations indexed by M. One way of finding the sub set M is to determine the sub set of size d, the deletion of which produces the largest reduction of the residual sum of squares. Problem is that d is rarely known, and finding (n_{c_d}) sub sets of size n-d with the minimum residual sum of squares involves extensive computations and some times can not be possible for large n. We suggest here to use robust ((*e.g.*, LMS, LTS (Rousseeuw 1984, 1985), RLS (Rousseeuw and Leroy, 1987)) and/or diagnostic techniques (*e.g.*, residual analysis, cook-type distances and/or robust distances) as a first aid for identifying suspected outliers and/or high-leverage points. Robust methods that are used for the identification of suspected group of observations can make the estimation robust with their own robustness properties, so we suggest giving the preference for using robust methods to find out initial suspicion-subset of unusual observations. We perform our plotting procedure by the following steps. First two steps for computations and last two steps help us to draw potential versus residual (P-R) plot.

5.4.2 Computation Steps

First step: Let us assume, d observations among a set of n observations are treated as unusual after the first diagnosis, a set of $(n\text{-}d)$ cases 'remaining' in the analysis is R and a set of cases which will be 'deleted' is D. Without loss of generality, we arrange these suspected cases at the end of data matrices of X, Y, and V is a variance-covariance matrix as follows.

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix} \quad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix} \quad V = \begin{bmatrix} V_R & 0 \\ 0 & V_D \end{bmatrix} \tag{5.17}$$

where $V_R = (X_R^T X_R)^{-1}$ and $V_D = (X_D^T X_D)^{-1}$.

Second step: In this stage we compute standardized deleted residuals by the formula

$$r^*_{st(R)} = \frac{r_{st(R)} - \bar{r}_{st(R)}}{S.E.(r_{st(R)})} \tag{5.18}$$

where $r_{st(R)} = Y - X\hat{\beta}_{(R)}$, $\hat{\beta}_{(R)} = (X_R^T X_R)^{-1} X_R^T Y_R$,

$\bar{r}_{st(R)} = mean(r_{st(R)})$ and $S.E.(r_{st(R)})$ is the residual standard error and to get the group-deleted potentials (name as potentials) we compute the diagonal elements of $H_{(R)}$,

$$h_{ii(R)} = x_i^T (X_R^T X_R)^{-1} x_i; \quad i = 1,2,...,n \tag{5.19}$$

and

$$H_{(R)} = X(X_R^T X_R)^{-1} X^T. \tag{5.20}$$

Third Step: We draw a scatter plot of potentials versus deleted standardized residuals. Since we transform the residuals into the standardized form (standard normal), we draw the 95% boundary lines (-1.96,1.96) in Y-axis for finding outliers and $median(h_{ii(R)}) + 3MAD(h_{ii(R)})$ cut-off line in X-axis for finding the high-leverage points. These types of cut of points for high-leverage points are suggested by Hadi (1992).

Forth Step: We draw a joint 95% confidence region (ellipsoid) by using deleted potentials and deleted standardized residuals according to the formula (idea) of

Mahalanobis distance for identifying influential observations, (see section 5.5). We consider the observation outside the confidence ellipse and in the region of outliers and/or high-leverage points as an influential observation. Since the equation of an ellipse is an additive form therefore we can say our influence statistic is far from multiplicative phenomena. Now the whole data set is classified into four as our requirements: regular (good) observations, outliers, high-leverage points and influential observations by making the figure fivefold. We say as Rousseeuw and van Zomeran (1990), "a single diagnostic can never be sufficient for this fivefold classification!"

5.4.3 Decisions

To illustrate the terminology and to reach in conclusion of our plotting procedure let us consider the figure 5.2, region (a), contains regular (good) observations inside the ellipse (or outside the regions of outliers and high-leverage points); region (b), assigns the observations possessing vertical outliers (regression outlier but not a leverage point), over the horizontal line at 1.96; region (c), contains observations outlying in x, y , below the horizontal line at -1.96; in region (d), right of the vertical cut-off line observations are treated as high-leverage points; and, for region (e), observations (outside the boundary of the 95% joint confidence region (ellipsoid) and at the same time in the regions of outliers and/or high-leverage points) are treated as influential observations.

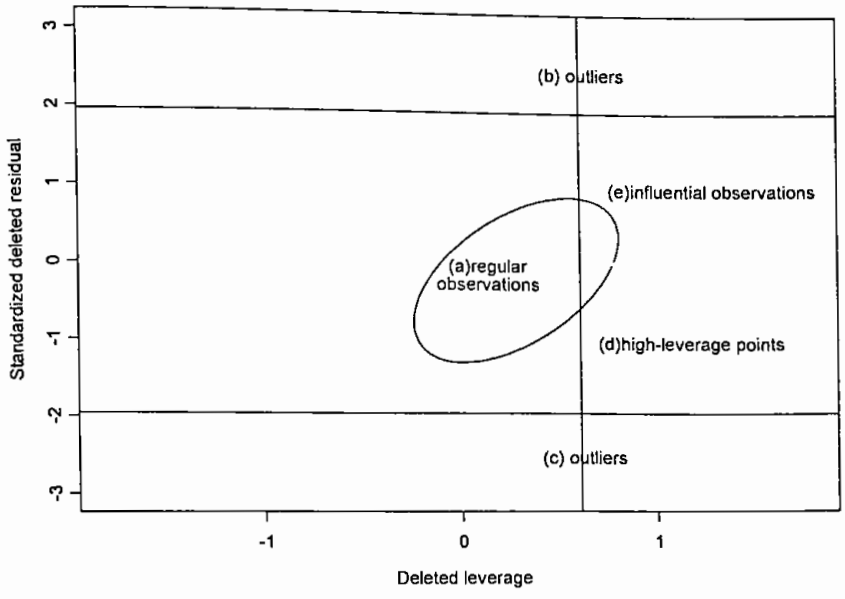


Figure 5.2 Classification of observations with (a) regular observations, (b) and (c) both are outliers, (d) high-leverage points, and (e) influential observations.

We assign the above classification numerically,

(i) $1.96 < outliers < -1.96$. (5.21)

(ii) High-leverage points > $median(h_{ii(R)}) + 3MAD(h_{ii(R)})$. (5.22)

(iii) Observations outside the confidence ellipsoid and at the same time lie into the regions of outliers' and/or high-leverage points are considered as influential observations. (5.23)

5.5 Proposed Distance

Now we want to formulate our proposed distance and find out the distribution of it. Following subsections are made to serve the purposes accordingly.

5.5.1 Derivation of Distance

It is well know that the Mahalanobis squared distance (*MSD*) based on the population mean and population scatter matrix is

$$MSD_i = (x_i - T(X))^T \Sigma^{-1} (x_i - T(X)), \tag{5.24}$$

where $X = (x_1, x_2, ..., x_n)$ is a data set matrix of n points in p dimensions and $T(X)$ is a vector of 'center' (mean) and Σ is the matrix of 'scatter' (covariance) of X.

Its (*MSD*) sample counterpart is defined as

$$MSD_i = (x_i - \overline{X})^T S^{-1} (x_i - \overline{X}), \tag{5.25}$$

where, $\overline{X}$ *and* S are the standard sample mean and covariance matrix.

We make a Mahalanobis type distance based on the sample mean $\overline{X}$ and sample scatter S of the matrix X_{PR} (potential-residual matrix), which is defined as

$$X_{PR}{}^T = (\textit{vector of potentials, vector of residuals}). \tag{5.26}$$

Since X (without deletion of suspects) may be contaminated by the unusual observations, the distance based on them will be sensitive to the unusual (extreme) observations. Becker and Gather (1999) pointed out, "classical tools based on the mean and covariance matrix are rarely able to detect entire multivariate outliers in given sample due to masking effect". To remedy from the problem of masking and/or swamping we try to make our distance resistant/robust by deleting the suspected group of unusual observations (extreme values) from the X. As a result we propose a type of robust distance

$$RDST_i = \sqrt{(x_i - T(X_{(R)}))^T S_{(R)}{}^{-1} (x_i - T(X_{(R)}))} \tag{5.26}$$

where $i = 1,2,...,n$ and

$$T(X_{(R)}) = \begin{bmatrix} mean & h_{ii(R)} \\ mean & r^*_{st(R)} \end{bmatrix} \text{ and } S_{(R)} = \text{var}\,iance(h_{ii(R)}, r^*_{st(R)}).$$

Since MD_i is the distance of the test point from the center of mass divided by the width of the ellipsoid in the direction of the test point, our robust distance (*RDST*) also measures

distance by drawing an ellipsoid that is constructed as an additive form after the deletion of the suspected group of unusual observations. Thus the ellipsoid (distance) holds robustness in itself.

5.5.2 Distribution of the Distance

The distribution of the Mahalanobis distance (*MD*) with the both true location and shape parameters is well known. For the better understanding we clarify the distributional results for the Mahalanobis type distance and to find the distribution of *RDST*. We consider three established distributional results.

1. If we consider the distance (d^2), *MSD* is based on true parameters μ and Σ and the data is normal then the d^2 follows χ^2 distribution (Mardia, Kent and Bibby, 1979),

$$i.e.\ d^2 = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \sim \chi^2_p . \tag{5.27}$$

2. If the distance (d^2), *MSD* is based on standard mean and covariance estimates this distance have an Beta distribution (Gnanadesikan and Kettenring 1972; Wilks, 1962),

$$i.e.\ \frac{(n-1)^2}{n} d^2 = \frac{(n-1)^2}{n} (x_i - \bar{X})^T S^{-1} (x_i - \bar{X}) \sim Beta\left(\frac{p}{n}, \frac{n-p-1}{2}\right), \tag{5.28}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $S = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{X})(x_i - \bar{X})^T$.

3. If the distance, d^2 is based on estimate, S of Σ that is independent of the x_i, S is an unbiased estimate of Σ based on sample of size n. these distance have an exact F distribution when μ is the location argument (Mardia, Kent and Bibbey, 1979) and an approximate F distribution when $\bar{X}$ is the location argument (Serfling, 1980). Given S and x_i are independent,

$$\frac{(n-p)}{(n-1)p} d^2 = \frac{(n-p)}{(n-1)p} (x_i - \mu)^T S^{-1} (x_i - \mu) \sim F_{\alpha(p,n-p)} . \tag{5.29}$$

Using a variant Slutsky's theorem

$$\frac{(n-p)}{(n-1)p}d^2 = \frac{(n-p)}{(n-1)p}\left(x_i - \bar{X}\right)^T S^{-1}\left(x_i - \bar{X}\right) \sim F_{\alpha(p,n-p)}. \qquad (5.30)$$

Hardin and Rocke in 2005 pointed out, "we apply an adjusted F distribution to the extreme sample points. The F distribution is more representative of the extreme points than the more commonly used χ^2 distribution". They showed in their paper that F distribution is also more appropriate than others in case of robust distance. If we consider the above arguments, we see our distance is analogous more appropriately to the 3rd (eq. 5.30) and we may come to the conclusion, *RDST* follows F distribution,

$$i.e.,\ RDST \sim \sqrt{\frac{(n-1)p}{(n-p)}F_{\alpha(p,n-p)}}\ . \qquad (5.31)$$

5.6 Examples

In this section we investigate the effectiveness of our proposed procedure empirically. To compare with existing identification techniques, we use three well-referred data sets from the literature and an artificial and high-dimensional large data set.

Hawkins et al. (1984) Data

We use the artificial data set generated by Hawkins *et al.* (1984) as our first example. It provides an example with masking effect. The data set consists of 75 observations in 4 dimensions, one for response variable others are explanatory variables. It is well established in the literature that the first 10 observations are bad leverage points and the next 4 observations are good leverage points. Hawkins *et al.* (1984) mentioned that the bad leverage points (outliers) are masked and the 4 good leverage points are appear outlying because they possess large residuals. Rousseeuw and Leroy (1987) showed that observations 1-10 are influential and the observations 11-14 (good leverage) are well accommodated by the LMS fit. Columns 2-5 in table 5.1 show single case diagnostics, Cooks distance identifies only observation 14 as influential , residuals r_i identify 5 observations (7, 11-14) leverage values identify only 3 (12-14) as extreme cases. *DFFITS* identifies observations 2, 7, 8,11,12,13 and14 as influential. Imon (2002) identified all the

observations (1-14) as high leverage points. Hardin and Rocke (2005) showed that all the 14 observations are masked when the mean and covariance are used to determine Mahalanobis squared distance. At this stage we apply the proposed algorithm to identify and classify the observations. We consider all first 14 observations as the suspected unusual observations and make the deletion set D by them. According to our algorithm table 5.1; by the columns 6 and 7, shows observations 1-10 as outliers and all 14 observations (1-14) are identified as high leverage points. Now, if we see to the figure 5.3 potential versus residual (P-R) plot, it clearly builds a confidence region (ellipse) by potentials and residuals in two dimensions. We know that first 10 observations are the most influential as they are at the same time, outliers and high leverage points. But these observations (1-10) are masked while the less influential (high leverage) 4 (11-14) points are getting more importance in DFFITS. Our P-R plot (figure 5.3) shows a very neat classification of 1-10 as outliers (good leverage), 11-14 as high leverage (bad leverage) points and all 14 observations as influential observations.

Table 5.1 Diagnostic measures for Hawkins *et al.* (1984) data

Index	r_i (2.50)	h_{ii} (0.107)	CD_i (1.00)	$\|DFFITS_i\|$ (0.462)	$\|r^*_{si(R)}\|$ (1.96)	$h_{ii(R)}$ (0.169)
1	1.55	0.063	0.040	0.406	**2.371**	**14.463**
2	1.83	0.060	0.053	**0.470**	**2.487**	**15.222**
3	1.40	0.086	0.046	0.430	**2.570**	**16.966**
4	1.19	0.086	0.031	0.352	**2.407**	**18.015**
5	1.41	0.081	0.039	0.399	**2.505**	**17.381**
6	1.59	0.073	0.052	0.459	**2.452**	**15.611**
7	**2.08**	0.068	0.079	**0.575**	**2.641**	**15.704**
8	1.76	0.063	0.052	**0.464**	**2.528**	**14.816**
9	1.26	0.080	0.034	0.372	**2.418**	**17.033**
10	1.41	0.087	0.048	0.439	**2.480**	**15.974**
11	**-3.66**	0.094	0.348	**-1.300**	0.185	**22.389**
12	**-4.50**	**0.144**	0.851	**-2.168**	0.176	**24.026**
13	**-2.88**	**0.109**	0.254	**-1.065**	0.351	**22.731**
14	**-2.56**	**0.564**	**2.114**	**-3.030**	0.202	**28.158**

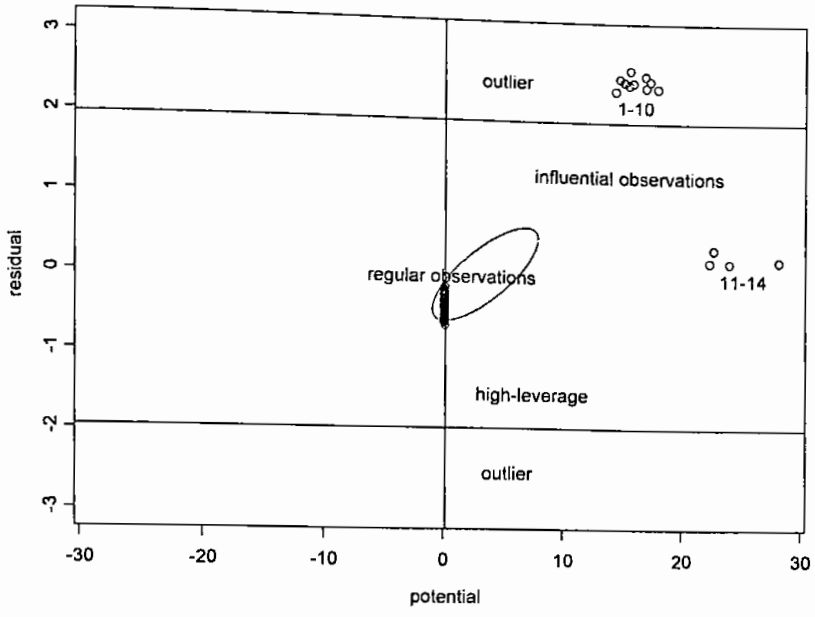


Figure 5.3 P-R plot; Classification of outliers, high-leverage points and influential observations for Hawkins *et al.* (1984) data

Stackloss Data

This data set comes from Brownlee (1965) and has an extensive use in regression diagnostics. It is a three-predictor data set with airflow, cooling water and acid concentration, consisting 21 observations. Rousseeuw and van Zomeren (1990) mentioned, 4 points (1, 2, 3 and 21) are regression outliers, observation 4 is vertical outlier, more over, they mentioned case 21 is not far from X-space and case 2 is a mild regression outlier. Hence we suspect all 5 observations (1, 2, 3, 4, and 21) are unusual and form the deletion group (D) by these 5. We calculate standardized deleted residuals and diagonal elements of deleted leverage (potential) matrix for the proposed method. Table 5.2 (columns 2,3) shows that LS residual predicts only case 21 as outlier and case 17 as high leverage point. Cook's distance does not able to identify any influential case and DFFITS can identify only 21 as influential observation. If we look at the last two

columns of the same table, we see our proposed method clearly identify observations 1, 2, 3, and 21 as high leverage points and observations 1 and 3 as outliers. When we look at our figure 5.4, we can easily and accurately classify 1, 3 and 21 are bad leverage, 4 is vertical outlier, and 4 observations (1, 2, 3 and 21) are most influential. According to the distance from the center of the data we can say observation 1 is the worst case in the whole data set.

Table 5.2 Diagnostic measures for Stack loss data

Index	$\|r_i\|$ **(2.50)**	h_{ii} **(0.381)**	CD_i **(1.00)**	$\|DFFITS_i\|$ **(0.873)**	$\|r^{\bullet}_{M(R)}\|$ **(1.960)**	$h_{ii\,(R)}$ **(0.609)**
1	1.19	0.302	0.154	0.795	**2.153**	**1.732**
2	-0.72	0.318	0.06	-0.481	0.956	**1.781**
3	1.55	0.175	0.126	0.744	**1.971**	**1.042**
4	1.89	0.129	0.131	0.788	1.938	0.263
5	-0.54	0.052	0.004	-0.125	-0.209	0.151
6	-0.97	0.077	0.020	-0.279	-0.332	0.188
7	-0.83	0.219	0.049	-0.438	-0.213	0.282
8	-0.48	0.219	0.017	-0.251	0.025	0.282
9	-1.05	0.140	0.045	-0.423	-1.051	0.188
10	0.44	0.200	0.012	0.213	-0.063	0.277
11	0.88	0.155	0.036	0.376	-0.059	0.199
12	0.97	0.217	0.065	0.509	-0.176	0.301
13	-0.48	0.158	0.011	-0.203	-0.780	0.223
14	-0.02	0.206	0.000	-0.009	-0.659	0.219
15	0.81	0.190	0.039	0.388	-0.265	0.283
16	0.30	0.131	0.003	0.113	-0.505	0.208
17	-0.61	**0.412**	0.065	-0.502	-0.394	0.512
18	-0.15	0.161	0.001	-0.065	-0.388	0.294
19	-0.20	0.175	0.002	-0.091	-0.274	0.282
20	0.45	0.08	0.004	0.131	0.239	0.102
21	**-2.64**	0.285	0.692	**-2.100**	-1.911	**0.849**

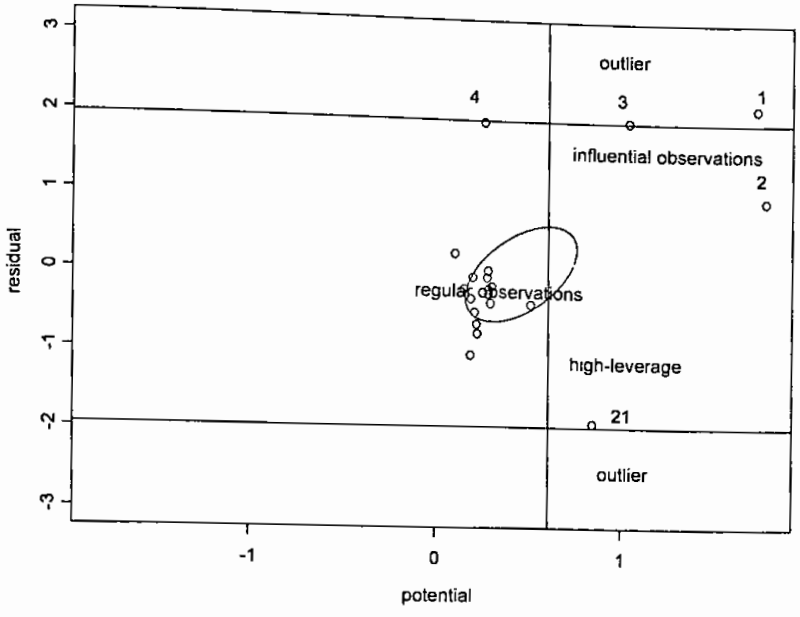


Figure 5.4 P-R plot; Classification of outliers, high-leverage points and influential observations for Stack loss data

Aircraft Data

Here we consider the well-known four-predictor Aircraft data that was presented by Gray in 1985. It deals with 23 single-engine aircraft built over the years 1947-1979. The dependent variable is cost, and the explanatory variables are aspect ratio, lift-to-drag ratio, weight of the plane, and maximal thrust. Most of the traditional diagnostic techniques identify case 22 as influential case. This observation possesses large Studentized residual, high-leverage values and huge Cook's distance and *DFFITS*. *DFFITS* also identifies observation 17 as influential, and observation 14 is identified as another high-leverage point. The most interesting feature is that the LMS and RLS technique fail to identify even a single observation as an outlier.

We now apply our proposed technique; we consider 3 observations (14, 17, and 22) as suspects by the existing popular methods. We find from the column 6 of table 5.3 that

observations 14, and 22 are outliers and column 7 shows (in the same table), observations 19, 21 and 22 as high-leverage points. However, when we draw the joint confidence region for finding influential observations we get all 4 observations (14, 19, 21 and 22) as influential observations. We finally reach the decision from the figure 5.5, observations 19 and 21 were masked and observation 17 was swamped before the diagnostic by our proposed method.

Table 5.3 Diagnostic measures for Aircraft data

Index	r_i **(2.50)**	h_{ii} **(0.435)**	CD_i **(1.00)**	$\|DFFITS_i\|$ **(0.933)**	$\|r^*_{st(R)}\|$ **(1.96)**	$h_{ii(R)}$ **(0.235)**
1	0.89	0.184	0.036	0.421	0.207	0.036
2	1.24	0.150	0.054	0.528	0.344	0.033
3	1.30	0.152	0.060	0.561	0.620	0.033
4	-0.73	0.156	0.020	-0.309	-0.296	0.036
5	-0.17	0.100	0.001	-0.056	-0.172	0.072
6	-0.94	0.257	0.061	-0.550	-0.278	0.179
7	-0.57	0.135	0.010	-0.221	-0.277	0.069
8	0.95	0.209	0.047	0.485	0.206	0.065
9	0.05	0.242	0.000	0.027	0.177	0.039
10	-0.97	0.218	0.052	-0.059	0.057	0.164
11	-0.29	0.166	0.003	-0.128	0.295	0.067
12	-1.83	0.062	0.044	-0.508	-0.906	0.090
13	-0.15	0.079	0.000	-0.044	-0.039	0.070
14	0.03	**0.880**	0.002	0.086	**-2.290**	0.122
15	-0.06	0.071	0.000	-0.016	-0.281	0.035
16	0.10	0.169	0.000	0.042	0.711	0.161
17	-1.83	0.243	0.215	**-1.117**	-1.131	0.166
18	-0.25	0.108	0.002	-0.085	0.184	0.107
19	0.66	0.283	0.035	0.410	0.256	**0.312**
20	1.08	0.153	0.042	0.462	-0.115	0.012
21	-0.49	0.309	0.022	-0.324	-0.398	**0.337**
22	**3.21**	**0.575**	**2.798**	**5.564**	**3.555**	**0.410**
23	-0.28	0.100	0.002	-0.092	-0.429	0.082

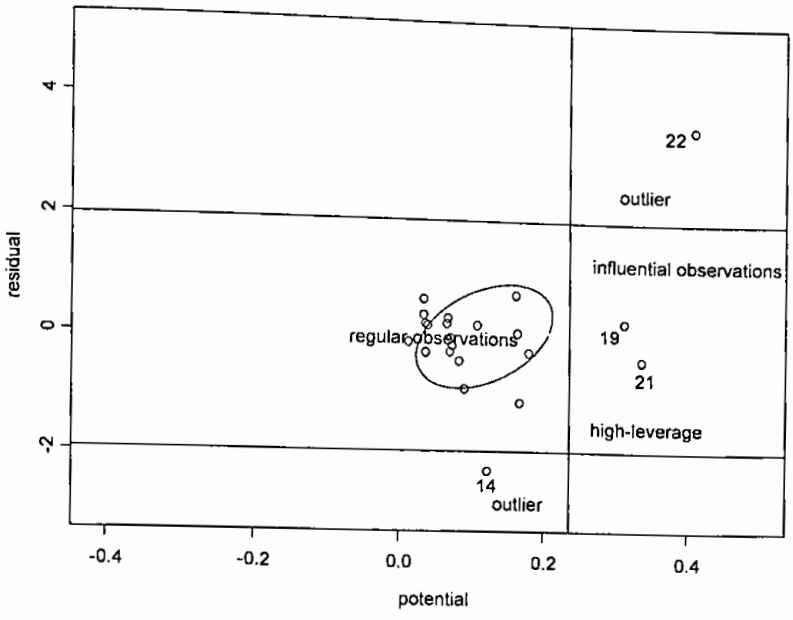


Figure 5.5 P-R plot; Classification of outliers, high-leverage points and influential observations for Aircraft data

High Dimensional and Large Artificial Data

Here we present an artificial large data set that is generated in a similar fashion described by Pena (2005). The data set is generated by the model

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_{10} x_{10} + \varepsilon . \tag{5.32}$$

We generate 500 observations, where X's have 11 dimensions with a constant term and they are independent random drawings from uniform distributions. The first 400 observations for each of the x_1-x_{10} variables are generated from Uniform (0, 10) and 100 observations (401-500) from Uniform (9, 10), makes the presence of heterogeneous variances in the data set. For the null model we generate errors from Normal (0, 1). The parameter values have been chosen as $\beta_0 = \beta_1 = ... = \beta_{10} = 1$, so that the standard diagnostics of the regression model does not show any evidence of heterogeneity.

Figure 5.6 shows our proposed plotting procedure has done the classification task well for large and high-dimensional data set as well as the previous examples.

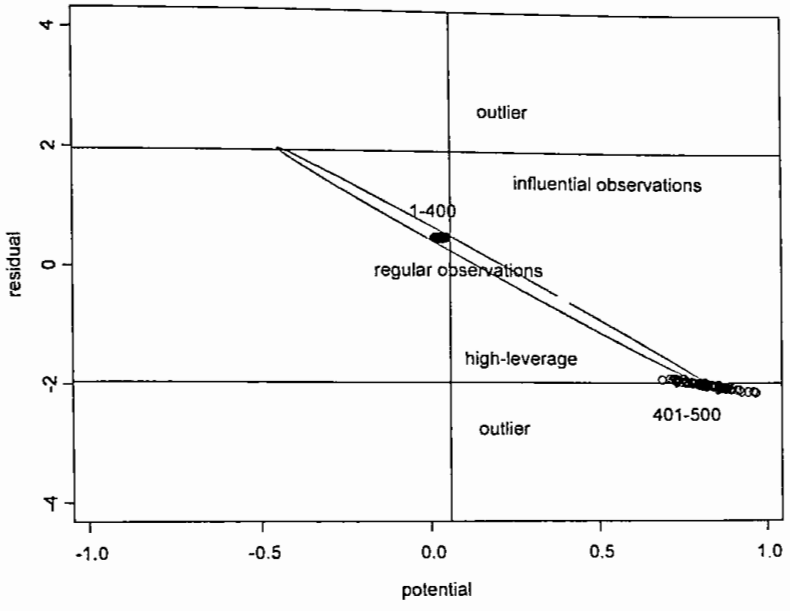


Figure 5.6 P-R plot; Classification of outliers, high-leverage points and influential observations for high dimensional, large and heterogeneous artificial data set.

Chapter 6

"One can not conceive anything so strange and so implausible that it has not already been said by one philosopher or another."
Rene descartes

Identification of Unusual Observations in Logistic Regression

Logistic regression diagnostics (LRD) have recently attracted much interest to the theoreticians as well as practitioners in recent years. It requires a higher mathematical level than most of the other material that steps backward to its study. This chapter presents different diagnostic aspects in logistic regression. As such linear regression, estimates of the logistic regression are sensitive to the unusual observations: outliers, high leverage and influential observations. Sections 6.1 and 6.2 cover the basic ideas of logistic regression and logistic regression diagnostics respectively. In section 6.3 we propose two new identification techniques for the multiple influential observations in binary logistic regression. The advantages and performance of the proposed methods in the identification of multiple influential cases are then investigated through several well-referred data sets in section 6.3.3.

6.1 Logistic Regression

Classically, logistic regression models were fit to data obtained under experimental conditions. The current use of logistic regression methods includes epidemiology, biomedical research, criminology, ecology, engineering, pattern recognition, wildlife biology, linguistics, business and finance *etc.* It is useful for situations in which we want

to be able to predict the presence or absence of a characteristics or outcome based on values of a set of predictor variables.

6.1.1 Difference between Linear and Logistic Regression

Main difference in a logistic regression model from the linear regression model is that the outcome variable is dichotomous. Logistic regression use one of three types of categorical response variables: binary, ordinal and nominal. Difference between logistic and linear regression is reflected both in the choice of a parametric model and in the assumptions. Among two most considerable differences, the first difference concerns the nature of the relationship between the outcome and explanatory variables. In linear regression the expected value of Y given the value of X, takes any value between -∞ and +∞. In logistic regression with dichotomous data, the conditional mean must be greater than or equal to zero and less than or equal to one. The second important difference between them concerns the conditional distribution of the outcome variable. For linear regression conditional distribution of the outcome variable Y given X will be normal with mean $E(Y|X)$, and a constant variance. Conditional distribution of the outcome variable of the logistic regression follows a binomial distribution with probability given by the conditional mean, $\pi(x)$. The changes in the conditional mean for per unit change in X becomes progressively smaller as the conditional mean gets closer to zero or one.

6.1.2 Logistic Regression Model Formulation

In any regression problem the key quantity is the mean value of the outcome variable, given the value of the explanatory variable(s), $E(Y/X)$. In linear regression we assume that this mean may be expressed as an equation linear in X (or some transformations of X or Y) such as

$$E(Y/X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{6.1}$$

hence

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \tag{6.2}$$

$$= X\beta + \varepsilon \tag{6.3}$$

$$= E(Y/X) + \varepsilon . \tag{6.4}$$

To represent the conditional mean of Y given X in case of logistic regression, we use the quantity $\pi(X) = E(Y/X)$. The specific relational form of the logistic regression model is

$$\pi(X) = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}} \quad ; 0 \le \pi(X) \le 1 \tag{6.5}$$

$$= \frac{\exp(Z)}{1 + \exp(Z)}, \tag{6.6}$$

where $Z = X\beta$. This form gives an S-curve configuration. The well-known 'logit' transformation in terms of $\pi(X)$ is

$$g(X) = \ln\left[\frac{\pi(X)}{1 - \pi(X)}\right] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p . \tag{6.7}$$

Hence, in logistic regression

$$Y = \pi(X) + \varepsilon , \tag{6.8}$$

here $\pi(x_i) = \pi_i$ is known as the probability for the i-th factor/covariate. Thus ε has a distribution with mean zero and variance equal to $\pi(X)[1 - \pi(X)]$.

6.1.3 Assumptions in Logistic Regression

Following assumptions are hold for performing logistic regression tasks:

1. The model is correctly specified, *i.e.,*
    a) The true conditional probabilities are a logistic function of the independent variables.
    b) No important variables are omitted.
    c) No extraneous variables are included, and
    d) The independent variables are measured without errors.
2. Cases are independent.

3. The independent variables are not linear combination of each other. Perfect multicolinearity makes estimation impossible, while strong multicolinearity makes estimation imprecise.

6.1.4 Parameter Estimation in Logistic Regression

We can use OLS for estimating parameters in logistic regression, but the assumptions under which the OLS estimators possesses very good properties, do not hold for logistic regression model. Mainly for this reason the maximum likelihood (ML) method based on iterative-reweighted least squares become the most popular with the statisticians.

If we let X denote the design matrix and Y denote the vector of response values, π denote the vector of $E(Y/X)$, the likelihood equation can be written as

$$\frac{\delta l\beta}{\delta\beta} = X^T(Y-\pi) \tag{6.9}$$

where $l(\beta) = \log(l(\beta)); and$ β is the parameter vector. From equation (6.9) it follows that

$$X^T\pi = X^TY, \tag{6.10}$$

since $\delta l / \delta\beta$ is set equal to zero. Since $\hat{Y} = \hat{\pi}$ (*i.e.*, the predicted value of Y_i is the estimated probability that $Y = 1$), the solution to equation (6.10) will satisfy

$$X^T\left(Y-\hat{Y}\right) = 0. \tag{6.11}$$

In linear regression, the likelihood equations, obtained by differentiating the sum of squared deviations function with respect to β are linear in the unknown parameters and thus are easily solved. But in logistic regression it is not possible for non linearity in β's and hence we use numerical optimization 'Newton-Raphson' method. This entails first determining $(\delta / \delta\beta)X^T(Y-\pi)$, which is equivalent to computing $-(\delta / \delta\beta)X^T\pi$, which equals $-[(\delta / \delta\beta)\pi]X$. From equation (6.5) we may obtain $\delta\pi / \delta\beta_0 = \pi(1-\pi)$ and $\delta\pi / \delta\beta_i = x_i\pi_i(1-\pi_i)$ .Hence we conclude $(\delta\pi / \delta\beta) = X^T\pi(1-\pi)$,

so $$-\frac{\delta}{\delta\beta}\left(X^T\pi\right) = -\frac{\delta\pi}{\delta\beta}X$$

$$\Rightarrow -X^T\pi(1-\pi)X \quad = -X^T WX\,, \tag{6.12}$$

where W is the diagonal matrix with elements $\pi_i(1-\pi_i)$, that would have to be estimated. Iterative estimate of β are then obtained as

$$\hat{\beta}_{i+1} = \hat{\beta}_i - \left[\frac{\delta^2 l\beta}{\delta\beta}\right]^{-1} \frac{\delta l\beta}{\delta\beta} \tag{6.13}$$

$$= \hat{\beta}_i + \left(X^T WX\right)^{-1} X^T (Y-\pi) \tag{6.14}$$

until estimated β converges to its previous value, *i.e.*, $\hat{\beta}_{i+1} = \hat{\beta}_i$.

6.1.5 Why Logistic Regression Diagnostics

Two main causes for logistic regression diagnostics that differ from linear regression are; MLE estimation method for parameter estimation in logistic regression is very sensitive to unusual observations and in presence of influential observations implicit assumption is broken down as like as in linear regression.

6.1.6 Notion of Outliers, High-Leverage Points and Influential Observations

In logistic regression the idea of outlier and influential point is somewhat different. In binomial logistic regression, we observe outlier or influential point (unusual observations) may occur mostly as: a) We see response variables (0, 1) are sometimes misclassified, b) By meaningful deviations (we see also low leverage) in explanatory variables, and c) Disagreement may come out in response and explanatory variables together. As a result all of the above three can break the normal pattern (S-curve) of the majority of the data.

6.1.7 The Basic Building Blocks of Logistic Regression Diagnostics

The role of a regression diagnostician is to provide routine methods of model sensitivity analysis which are both intuitively appealing and inexpensive. Clearly this requires a through understanding of the model and the nature of the fitting process.

Pregibon mentioned in 1981, "For The logistic regression, the basic building blocks for the identification of outlying and influential points will again be a residual vector and a projection matrix". In logistic regression model, residuals can be defined on several scales. The two most useful are the components of chi-square (Pearson's residual) and the components of deviance. Hosmer and lemeshow (2000) pointed out that the key quantities for the logistic regression diagnostics, as in linear regression, are the components of the 'residual sum-of-squares' and the deviance for logistic regression. Both of them play the same role that the residual sum of squares plays in linear regression.

6.2 Logistic Regression Diagnostics

We want to detect the influential observations to correct and/or remove them from the data set that makes the decision more meaningful and achieves analysts' insight.

A large body of literature is available (see Belsley *et al.*, 1980; Rousseeuw and Leroy, 1987; Chatterjee and Hadi, 1988; Cook and Weisberg, 1982; Ryan, 1997; Hosmer and Lemeshow, 2000) for the identification of unusual points. Pregibon (1981) provided the theoretical work that extended linear regression diagnostics to logistic regression.

As the linear regression, identification of unusual observations in logistic regression are classified into three categories:

1. Outliers identification,
2. High leverage points identification, and
3. Influential observations identification.

Besides the formal diagnostic procedures, a number of different types of plotting procedures have been suggested for use of diagnostics in logistic regression. These consist of the following:

(a) $Plot\ \Delta X_j^2 \quad versus \quad \hat{\pi}_j$

(b) $Plot\ \Delta D_j \quad versus \quad \hat{\pi}_j$

(d) $Plot\ \Delta X_j^2 \quad versus \quad h_{jj}$

(e) $Plot\ \Delta D_j \quad versus \quad h_{jj}$,

(c) $Plot\ \Delta\hat{\beta}_j$ *versus* $\hat{\pi}_j$ (f) $Plot\ \Delta\hat{\beta}_j$ *versus* h_{jj}

where ΔX_j^2, ΔD_j *and* $\Delta\hat{\beta}_j$ are the Pearson chi-square statistic, change in the deviance and change in the estimated parameters respectively.

This section presents a short discussion of different diagnostic measures for logistic regression that are originated from linear regression. We also try to modify and develop the most popular diagnostic measures for logistic regression based on the work of Pregibon (1981).

6.2.1 Identification of Outliers

Most of the times outliers are identified by using residuals or some functions of the residual. In logistic regression, the *i*-th residual is defined as

$$\hat{\varepsilon}_i = y_i - \hat{\pi}_i, \quad i = 1,2,...,n. \tag{6.15}$$

Residuals measure the extent of ill-fitted factor/covariate patterns. Sometimes we also suppose that there are j distinct values of observed x. We denote the number of cases $x = x_j$ *by* m_j, $j = 1,2,...,j$. In this situation we define the j-th residual as

$$\hat{\varepsilon}_j = y_j - m_j\hat{\pi}_j, \quad j = 1,2,...,n. \tag{6.16}$$

We assume for the simplicity in our study, m_j =1. The observations possessing large residuals are suspect outliers. The unscaled residuals are not readily applicable in detecting outliers. Now let us introduce with some scaled version of residuals that are commonly used in diagnostics for the identification of outliers. In logistic regression the error variance is a function of the conditional mean, *i.e.*,

$$Var(y_i / x_i) = v_i = m_i\hat{\pi}_i(1-\hat{\pi}_i). \tag{6.17}$$

The Pearson residual defined for the *i*-th factor/covariate pattern is given by

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i}}, \quad i = 1, 2, ..., n. \tag{6.18}$$

Pregibon (1981) derived a linear approximation to the fitted values, which yields a hat matrix for logistic regression

$$H = V^{1/2}X(X^TVX)^{-1}X^TV^{1/2}, \tag{6.19}$$

where V is the diagonal matrix with general elements v_i.

The above approximation yields

$$\hat{\varepsilon}_i = y_i - \hat{\pi}_i \approx (1 - h_{ii})y_i \tag{6.20}$$

and the variance of the residual is given by

$$V(\hat{\varepsilon}_i) = v_i(1 - h_{ii}) \tag{6.21}$$

which suggests that the Pearson residuals do not have variance equal to 1. For this reason we could use the standardized Pearson residual given by

$$r_{si} = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i(1 - h_{ii})}}, \quad i = 1,2,...,n. \tag{6.22}$$

"The quantities r_i, r_{si} and h_{ii} (i-th diagonal element of leverage/projection matrix) are useful for detecting extreme points, but not for assessing their impact on the various aspect of the fit" (Pregibon 1981). Standardized Pearson's residual is suggested to use of single outlier identification and the i-th observation is termed as outlier if $|r_{si}| > 3$ (equation 6.22). Draper and John (1981) pointed out that observation with the largest residual was not the most influential, however, deletion of observation possessing a small residual, had a marked effect on the parameter estimates. But the reality is no guarantee that the data set will contain just a single outlier. Hampel *et al.* (1986) pointed, "A routine data set may contain about 10% outliers in it". A group of outliers may cause of masking and swamping and as a result distort the fitting of a model in such a way that outliers may have artificially very small residuals so that they may appear as inliers. A number of diagnostic procedures have been suggested to identify multiple outliers in linear regression, but so far as we know this issue is not much addressed in logistic regression. We mention here the generalized standardized Pearson residual (GSPR) suggested by Imon and Hadi (2005), defined as

$$r_{si(R)} = \begin{cases} \dfrac{y_i - \hat{\pi}_{i(R)}}{v_{i(R-i)}(1-h_{ii(R)})} \; for \quad i \in R \\ \dfrac{y_i - \hat{\pi}_{i(R)}}{v_{i(R)}(1+h_{ii(R)})} \; for \quad i \notin R \end{cases} \tag{6.23}$$

where

$$h_{ii(R)} = \hat{\pi}_{i(R)}(1-\hat{\pi}_{i(R)}) x_i^T \left(X_R^T V_R X_R\right)^{-1} x_i \tag{6.24}$$

$$\hat{v}_{i(R)} = \hat{\pi}_{i(R)}(1-\hat{\pi}_{i(R)}) \; ; \tag{6.25}$$

a group of observations D is omitted, and the fitted values for the entire logistic regression model based on R (group of remaining observations) set are defined as

$$\hat{\pi}_{i(R)} = \frac{\exp\left(x_i^T \hat{\beta}_{(R)}\right)}{1+\exp\left(x_i^T \hat{\beta}_{(R)}\right)}, \qquad i = 1,2,...,n. \tag{6.26}$$

6.2.2 Identification of High Leverage Points

In 1981 Pregibon derived the diagonal of the leverage and projection matrix for logistic regression as

$$h_{ii} = m_i \hat{\pi}_i (1-\hat{\pi}_i) x_i^T (X^T V X)^{-1} x_i \tag{6.27}$$

and m_i , which is the i-th diagonal element of $M=(I-H)$ respectively. The i-th observation bearing large h_{ii} and small m_i are used to treat as an extreme case in the design space.

6.2.3 Identification of Influential Observations

Among the diagnostic statistics in linear regression, Cook's distance (Cook 1977, 1979), *DFFITS*, and *DFBETA* (Belsley *et al.*, 1980) have become very popular with the practitioners of logistic regression. We may define the above statistics for logistic regression as the following way. We define the i-th Cook's distance as

$$CD_i = \frac{\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)^T (X^T V X)\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)}{k v_i}; \quad i = 1,2,...,n, \tag{6.28}$$

where $\hat{\beta}^{(-i)}$ is the estimated parameter of β with the i-th observation deleted. The i-th Cook's distance can be re-expressed in terms of the i-th standardized Pearson residual and leverage as

$$CD_i = \frac{1}{k} r_{si}^2 \left(\frac{h_{ii}}{1 - h_{ii}} \right) \tag{6.29}$$

We call an observation influential if its corresponding CD_i value is greater than 1.

We may define i-th DFBETA as Besley *et al.* (1980),

$$DFBETA_i = \hat{\beta}_i - \hat{\beta}_i^{(-i)} \tag{6.30}$$

and re-expressed by leverage and residual as

$$DFBETA_i = \frac{(X^T V X)^{-1} x_i^T \hat{\varepsilon}_i}{1 - h_{ii}} \quad . \tag{6.31}$$

DFFITS may be defined for logistic regression as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{v_i^{(-i)} \sqrt{(1 - h_{ii})}}; \quad i = 1,2,...,n \tag{6.32}$$

where $\hat{y}_i^{(-i)}$ and $v_i^{(-i)}$ are respectively the i-th fitted response and the estimated standard error with the i-th observation deleted. *DFFITS* values can be expressed in terms of standardized Pearson residual and leverage value as

$$DFFITS_i = r_{si} \sqrt{\frac{h_{ii}}{(1 - h_{ii})} \frac{v_i}{v_i^{(-i)}}} . \quad i = 1,2,...,n \tag{6.33}$$

Observation possessing *DFFITS* value grater than $3\sqrt{k/n}$ is termed as an influential observation.

We also introduce our newly proposed method for logistic regression as

$$SDFBETA_i = \frac{(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)^T (X^T V X)(\hat{\beta}_i^{(-i)} - \hat{\beta}_i)}{v_i^{(-i)}(1 - h_{ii})} \tag{6.34}$$

where $\hat{\beta}_i^{(-i)}$ is the estimated parameter vector and $v_i^{(-i)}(1 - h_{ii})$ is the variance of the i-th residual respectively after deleting i-th observation.

We make similar relationship of *CD*, *DFFITS* and *SDFBETA*,

$$SDFBETA_i = \frac{CD_i p v_i}{v_i^{(-i)}(1 - h_{ii})} \tag{6.35}$$

$$= \frac{DFFITS_i^2}{1-h_{ii}}. \tag{6.36}$$

Cut-off value: We may consider the i-th observation to be influential if it satisfies the condition

$$|\text{SDFBETA}_i| \geq \frac{(3\sqrt{k/n})^2}{1-(3p/n)} = \frac{9k}{n-3p}; \text{ where } k = p+1. \tag{6.37}$$

6.2.4 Examples

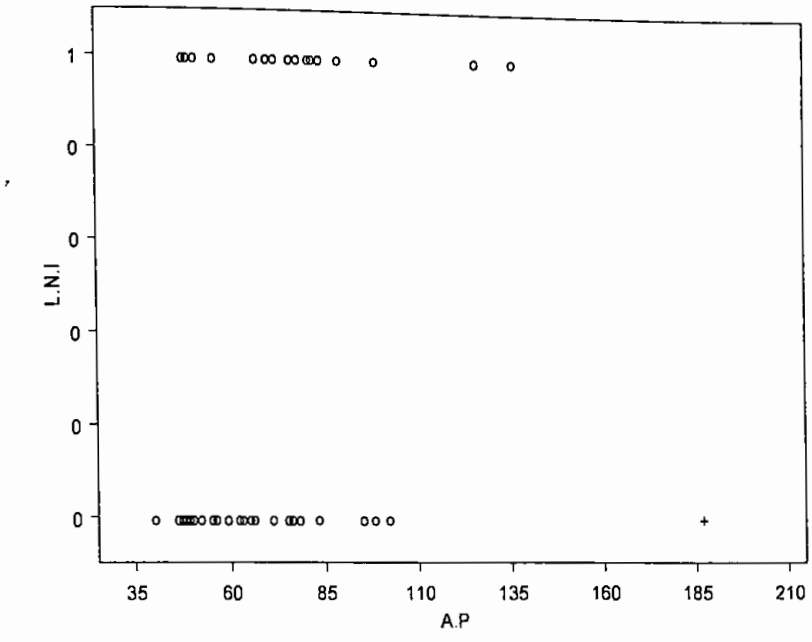
To show the performance of the proposed SDFBETA with single deletion diagnostic measures: Cook's distance (CD) and DFFITS, we consider first the well-known Brown data and then we make a modification in the data and show the comparison.

Brown Data

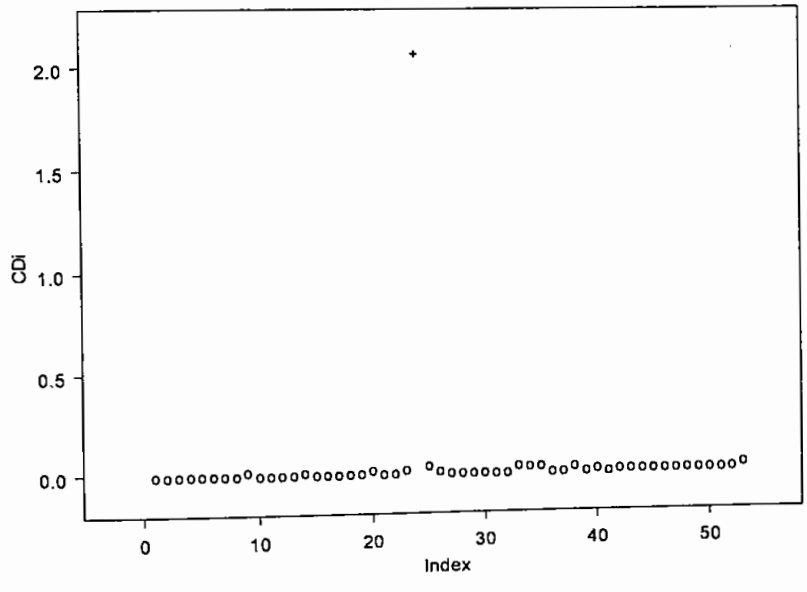
To illustrate the performance of identification task of the proposed method squared difference in BETA (SDFBETA) we consider part of a data set in Brown et al., (1980). The original objective was to see whether an elevated level of acid phosphates (A.P.) in the blood serum would be of value as an additional regressor for predicting whether or not prostate cancer patients also had lymph node involvement (L.N.I). The data set in Brown *et al.* (1980) additionally contains data on the four more commonly used regressors, but we use only acid phosphates in illustrating simple logistic regression. The data on the 53 patients are given in table (A.6) in Appendix; the dependent variable is nodal involvement, with 1 denoting the presence of nodal involvement, and 0 indicating the absence of such involvement. Scatter plot, figure 6.1 (a) shows there is clearly an unusual observation (187, case 24) among the patients without nodal involvement. Ryan (1997) considers the 24th observation as an outlier. We apply our new diagnostic measure for the identification of the influential case. Table 6.1 shows the evidence that SDFBETA along with Cook's distance and DFFITS successfully identify the case 24 as an influential. Figure 6.1 (b,c,d) justify the identification in favor of the single case deletion diagnostic measures.

Table 6.1 Diagnostic measures for Brown data; one outlier

Index	*CDi* (1.000)	*|DFFITSi|* (0.582)	*|SDFBETA|* (0.353)	Index	*CDi* (1.000)	*|DFFITSi|* (0.582)	*|SDFBETA|* (0.353)
1	0.006	-0.107	-0.110	28	0.006	0.106	0.109
2	0.005	-0.104	-0.107	29	0.006	-0.106	-0.109
3	0.006	-0.106	-0.109	30	0.006	-0.110	-0.115
4	0.006	-0.105	-0.108	31	0.005	-0.104	-0.107
5	0.006	-0.106	-0.109	32	0.005	-0.104	-0.106
6	0.006	-0.106	-0.110	33	0.038	0.280	0.289
7	0.006	-0.108	-0.112	34	0.033	0.259	0.266
8	0.005	-0.104	-0.106	35	0.036	0.272	0.281
9	0.025	0.228	0.234	36	0.006	-0.107	-0.110
10	0.005	-0.104	-0.107	37	0.006	-0.104	-0.107
11	0.005	-0.104	-0.106	38	0.032	-0.254	-0.267
12	0.006	-0.112	-0.114	39	0.008	-0.123	-0.125
13	0.006	-0.105	-0.108	40	0.021	-0.206	-0.214
14	0.017	0.187	0.191	41	0.006	-0.106	-0.108
15	0.006	-0.107	-0.111	42	0.018	0.188	0.193
16	0.006	-0.106	-0.110	43	0.017	0.184	0.188
17	0.006	-0.106	-0.109	44	0.016	0.180	0.184
18	0.008	-0.128	-0.131	45	0.016	0.182	0.186
19	0.011	-0.145	-0.149	46	0.016	0.181	0.185
20	0.025	-0.226	-0.236	47	0.016	0.182	0.186
21	0.006	-0.105	-0.108	48	0.017	0.187	0.191
22	0.007	-0.120	-0.122	49	0.017	0.185	0.190
23	0.025	0.222	0.232	50	0.017	0.187	0.191
24	**2.075**	**-2.149**	**-3.619**	51	0.016	0.181	0.184
25	0.044	0.294	0.343	52	0.020	0.198	0.204
26	0.017	0.185	0.190	53	0.040	0.281	0.316
27	0.006	-0.110	-0.115				

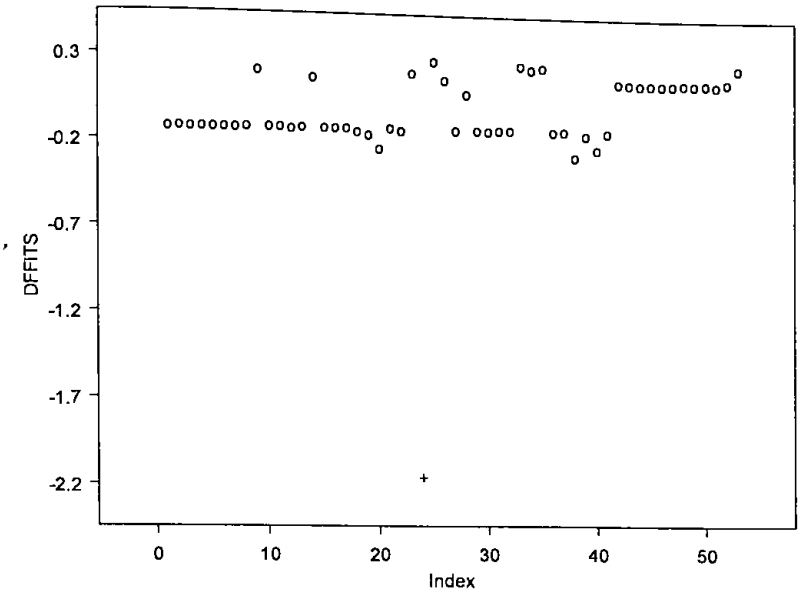


(a)

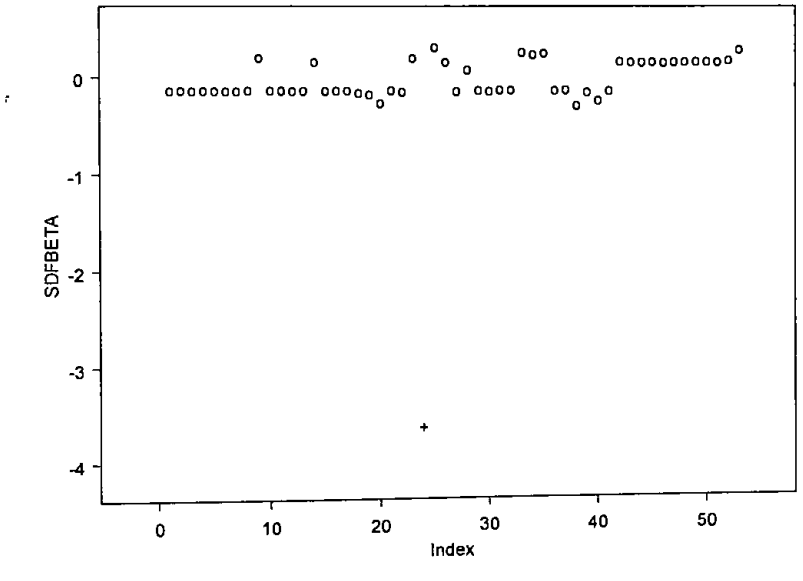


(b)

Figure 6.1 (a) Scatter plot of acid phosphates versus nodal involvement, (b) Index plot of Cook's distance



(c)



(d)

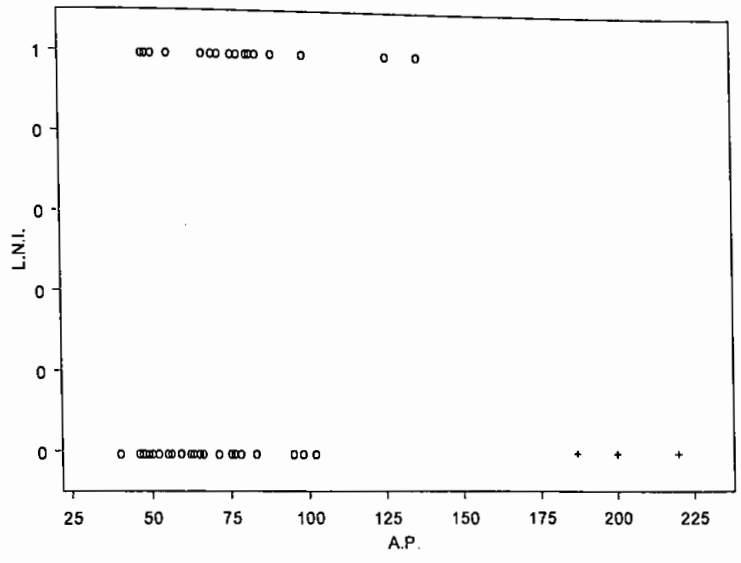
Figure 6.1 (c) Index plot of DFFITS, and (d) Index plot of SDFBETA

Modified Brown data

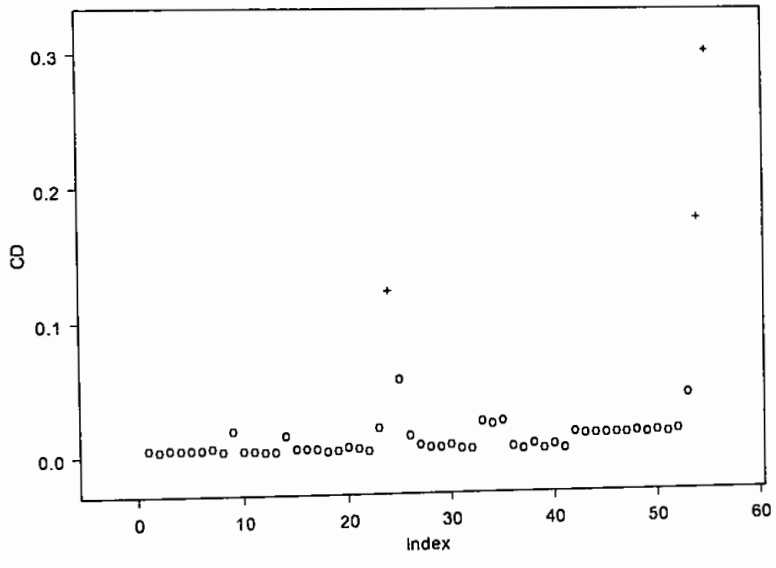
Here we modify the Brown (1980) data by putting two more unusual observations as cases 54 and 55 and this data set is presented in table (A.7) in the appendix. Scatter plot of figure 6.2 (a) shows the isolation of these points. Now we apply above three diagnostic measures again. Cook's distance and *DFFITS* do not identify the unusual observations properly. *DFFITS* identifies only 2 observations (54, 55), but *SDFBETA* identifies 3 observations (24, 54 and 55) correctly but at the same time swamps one (case 25) more, and *CD* is totally failed to identify the influential cases. That is, none of them are reliable for the identification of the multiple influential cases. Though the figures 6.2 (b,c,d), we show the separation of the cases regular and unusual but table 6.2 (by using cut-off values) shows the clear indication of failure of the measures.

Table 6.2 Diagnostics measures for modified Brown data; 3 outliers

Index	|DFFITS| (0.572)	CD (1.000)	|SDFBETA| (0.340)	Index	|DFFITS| (0.572)	CD (1.000)	|SDFBETA| (0.340)
1	-0.119	0.007	-0.123	29	-0.117	0.007	-0.120
2	-0.110	0.006	-0.112	30	-0.131	0.009	-0.136
3	-0.117	0.007	-0.120	31	-0.111	0.006	-0.113
4	-0.114	0.007	-0.117	32	-0.107	0.006	-0.109
5	-0.117	0.007	-0.120	33	0.232	0.026	0.239
6	-0.118	0.007	-0.121	34	0.222	0.024	0.228
7	-0.122	0.008	-0.126	35	0.229	0.026	0.235
8	-0.105	0.006	-0.107	36	-0.119	0.007	-0.123
9	0.207	0.021	0.212	37	-0.104	0.005	-0.106
10	-0.111	0.006	-0.113	38	-0.136	0.009	-0.140
11	-0.105	0.006	-0.107	39	-0.102	0.005	-0.104
12	-0.101	0.005	-0.103	40	-0.122	0.008	-0.125
13	-0.103	0.005	-0.105	41	-0.103	0.005	-0.105
14	0.186	0.017	0.189	42	0.184	0.017	0.187
15	-0.121	0.007	-0.124	43	0.181	0.016	0.185
16	-0.118	0.007	-0.121	44	0.180	0.016	0.183
17	-0.117	0.007	-0.120	45	0.182	0.016	0.186
18	-0.103	0.005	-0.105	46	0.180	0.016	0.183
19	-0.106	0.006	-0.109	47	0.182	0.016	0.186
20	-0.128	0.008	-0.131	48	0.186	0.017	0.189
21	-0.114	0.007	-0.117	49	0.182	0.016	0.186
22	-0.102	0.005	-0.104	50	0.186	0.017	0.189
23	0.211	0.022	0.217	51	0.181	0.016	0.184
24	-0.498	0.124	**-0.614**	52	0.190	0.018	0.194
25	0.341	0.058	**0.366**	53	0.300	0.045	0.318
26	0.182	0.016	0.186	**54**	**-0.594**	0.176	**-0.773**
27	-0.131	0.009	-0.136	**55**	**-0.782**	0.303	**-1.123**
28	-0.117	0.007	-0.120				

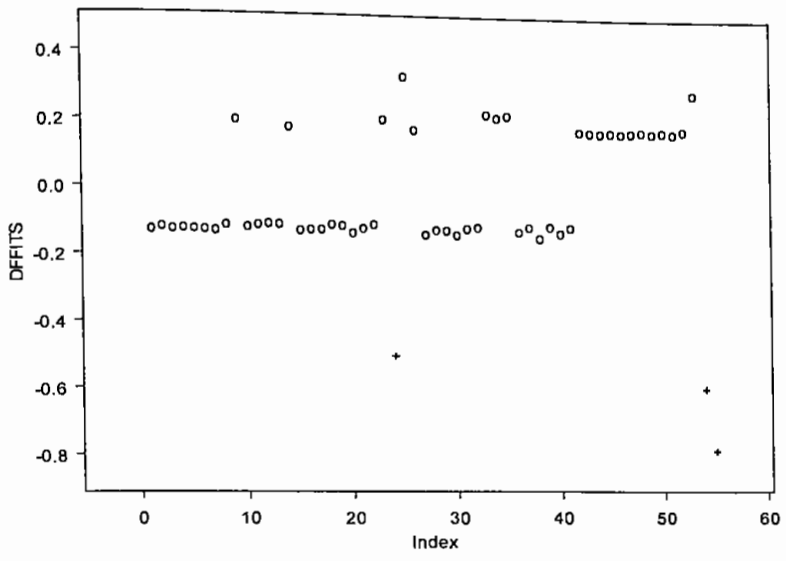


(a)

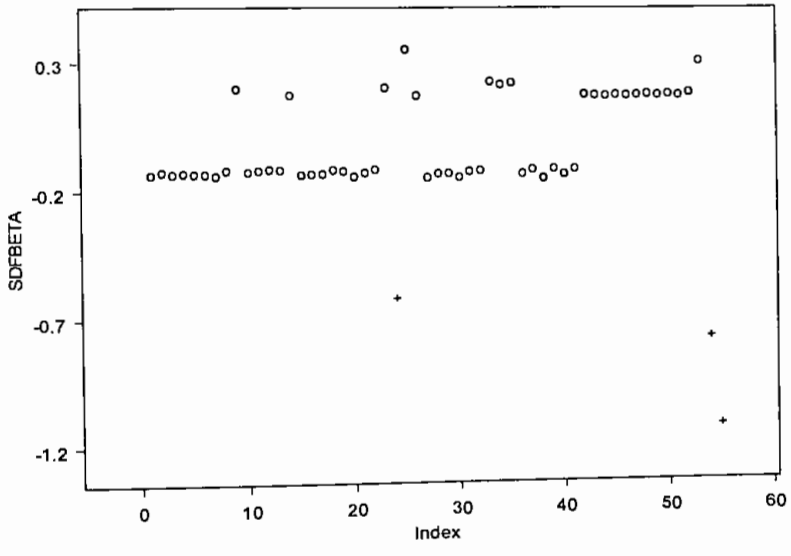


(b)

Figure 6.2 (a) Scatter plot of acid phosphates (A.P.) versus nodal involvement (L.N.I.); (b) Index plot of Cook's distance (in presence of 3 unusual observations)



(c)



(d)

Figure 6.2 (c) Index plot of DFFITS; and (d) Index plot of SDFBETA

(In presence of 3 unusual observations)

6.3 Identification of Multiple Influential Observations

In this section we propose two group deleted version of diagnostic measures for identifying multiple influential observations in logistic regression. These proposed measures are very much similar to the group deletion idea of Hadi and Simonoff (1993) and Atkinson (1994).

As noted earlier, a general approach of unusual observations identification is to form a clean subset of data and a subset of suspect group of unusual (outliers/high leverage points) observations, and then test the sensitivity of the observations on the estimated parameters and on the model analysis before and after the deletion of suspect group of cases. This is essential to find out the suspect group of d cases with the minimum residual sum of squares. Problem with this approach is that d is rarely known and some times even impossible to compute, it mostly depends upon the value of d comparing with total number of observations (n). Some times graphical displays like scatter plot, index plot and character plot (two or three regressors and a response variable) of explanatory and response variables may give us an idea about the suspect group of unusual observations, but these plots are not helpful for higher dimension of regressors. Like many others Atkinson (1986), Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990) suggest to use robust regression and/or diagnostic techniques to find the suspects. As Rousseeuw and Leroy (1987), we suggest to use robust techniques for overcoming the problem of masking/swamping and it is now evident that most of the times results are fruitful to identify outliers in presence of a number of unusual observations. Here one can use robust regression techniques like LMS (Rousseeuw, 1984), LTS (Rousseeuw, 1985), and re-weighted least squares (RLS) (Rousseeuw and Leroy 1987) for finding the suspect unusual observations.

We assume that d observations among a set of n observations are identified as suspect cases. Let us denote a set of cases 'remaining' in the analysis by R and a set of cases 'deleted' by D. Hence R contains (n-d) cases after d cases (group D) are deleted. Without loss of generality, we assume that these observations are the last of d rows of X, Y and V (variance-covariance matrix) so that we make

$$X=\begin{bmatrix} X_R \\ X_D \end{bmatrix} \qquad Y=\begin{bmatrix} Y_R \\ Y_D \end{bmatrix} \qquad V=\begin{bmatrix} V_R & 0 \\ 0 & V_D \end{bmatrix}.$$

Now we compute group-deleted version of the residuals and weights (leverage values) that will use later to develop the new diagnostic measures for the identification of multiple influential observations in logistic regression. Necessary results from the above arrangements are

$$\hat{\varepsilon}_{i(R)} = y_i - \hat{\pi}_{i(R)} \text{ and } \quad h_{ii(R)} = \hat{\pi}_{i(R)}(1-\hat{\pi}_{i(R)})x_i^T (X_R^T V_R X_R)^{-1} x_i ,$$

where $\hat{\pi}_{i(R)}$ is the estimated value for the i-th case after the deletion of d suspect cases, *i.e.*, based on the remaining group of observations R.

6.3.1 Generalized DFFITS (GDFFITS)

In this sub section we would like to introduce a generalized version of *DFFITS*, designed for logistic regression model, we name this *GDFFITS*. When a group of observations D is omitted, the fitted values for the entire logistic regression model based on R set are defined as

$$\hat{\pi}_{i(R)} = \frac{\exp\left(x_i^T \hat{\beta}_{(R)}\right)}{1+\exp\left(x_i^T \hat{\beta}_{(R)}\right)}; \quad i = 1,2,...,n .$$

We also define the i-th residual variance and the i-th diagonal element of the leverage matrix as

$$\hat{v}_{i(R)} = \hat{\pi}_{i(R)}(1-\hat{\pi}_{i(R)})$$

and $$h_{ii(R)} = \hat{\pi}_{i(R)}(1-\hat{\pi}_{i(R)})x_i^T \left(X_R^T V_R X_R\right)^{-1} x_i \qquad \text{respectively.}$$

Using the above results and also using linear-regression like approximation, we define generalized *DFFITS* for the logistic regression model as

$$GDFFITS_i = \begin{cases} \dfrac{\hat{y}_{i(R)} - \hat{y}_{i(R-i)}}{\sqrt{v_{i(R-i)}\, h_{ii(R)}}} \; for \quad i \in R \\ \dfrac{\hat{y}_{i(R+i)} - \hat{y}_{i(R)}}{\sqrt{v_{i(R)}\, h_{ii(R+i)}}} \; for \quad i \notin R \end{cases}. \tag{6.38}$$

Results are

$$\hat{y}_{i(R+i)} = \hat{y}_{i(R)} + \frac{h_{ii(R)}}{1+h_{ii(R)}}\hat{\varepsilon}_{i(R)} \ , \tag{6.39}$$

$$\hat{\beta}_{i(R+i)} = \hat{\beta}_{i(R)} + \frac{(X_R^T V_R X_R)^{-1}}{1+h_{ii(R)}}\hat{\varepsilon}_{i(R)} \tag{6.40}$$

and

$$h_{ii(R+i)} = \frac{h_{ii(R)}}{1+h_{ii(R)}}, \tag{6.41}$$

help us to re-express the *GDFFITS* quantities in terms of *GSPR* and deleted leverages as

$$GDFFITS_i = r_{si(R)}\sqrt{h_{ii}^*}\ ; \quad i = 1,2,...,n \tag{6.42}$$

where

$$h_{ii}^* = \begin{cases} \dfrac{h_{ii(R)}}{1-h_{ii(R)}} \ for \quad i \in R \\ \dfrac{h_{ii(R)}}{1+h_{ii(R)}} \ for \quad i \notin R \end{cases} . \tag{6.43}$$

The detection rule of influential observations suggested for *GDFFITS* in linear regression (as Imon 2005) may apply for *GDFFITS* in logistic regression. We consider *i*-th observation as influential if

$$\left|GDFFITS_i\right| \geq 3\sqrt{k/(n-d)} \ . \tag{6.44}$$

6.3.2 Generalized Squared Difference in Beta (GSDFBETA)

In a similar fashion of linear regression we suggest a generalized version of the squared difference in beta (*GSDFBETA*) for the identification of multiple influential observations in logistic regression.

We can define the generalized squared difference in beta (GSDFBETA) for the entire data set, in presence of a group of influential observations, as

$$GSDFBETA_i = \begin{cases} \dfrac{\left(\hat{\beta}_{(R)} - \hat{\beta}_{(R-i)}\right)^T (X_R{}^T V_R X_R)\left(\hat{\beta}_{(R)} - \hat{\beta}_{(R-i)}\right)}{v_{i(R-i)}\left(1-h_{ii(R)}\right)}; & i \in R \\ \dfrac{\left(\hat{\beta}_{(R+i)} - \hat{\beta}_{(R)}\right)^T (X_D{}^T V_D X_D)\left(\hat{\beta}_{(R+i)} - \hat{\beta}_{(R)}\right)}{v_{i(R)}\left(1-h_{ii(R+i)}\right)}; & i \notin R \end{cases} \tag{6.45}$$

It also can re-express in terms of generalized standardized Pearson residuals (*GSPR*) and deleted leverages h_{ii}^{*} as

$$GSDFBETA_i = \begin{cases} \dfrac{h_{ii}^{*} r_{si(R)}^2}{1-h_{ii(R)}} \; for \quad i \in R \\ \dfrac{h_{ii}^{*} r_{si(R)}^2}{1+h_{ii(R)}} \; for \quad i \notin R \end{cases}. \tag{6.46}$$

Using the relationship of *GSDFBETA* with other diagnostic *GDFFITS* as given in linear regression, we get the relation for logistic regression as follows:

$$GSDFBETA_i = \begin{cases} \dfrac{GDFFITS_i^{\,2}}{1-h_{ii(R)}}; \quad i \in R \\ \dfrac{GDFFITS_i^{2}}{1+h_{ii(R)}}; \quad i \notin R \end{cases}. \tag{6.47}$$

Cut-off value: Observations corresponding to large GSDFBETA comparing with maximum regular observations are declared as influential observations. Since the theoretical distribution of GSDFBETA is not so easy we should make a boundary value type cut-off for them. We may consider the *ith* observation to be influential in logistic regression as like as linear regression of *GSDFBETA*, if it satisfies the condition,

$$\begin{aligned} |GSDFBETA| &\geq \frac{(3\sqrt{k/(n-d)})^2}{1-[3p/(n-d)]} \\ &= \frac{9k}{n-d-3p}. \end{aligned} \tag{6.48}$$

6.3.3 Examples

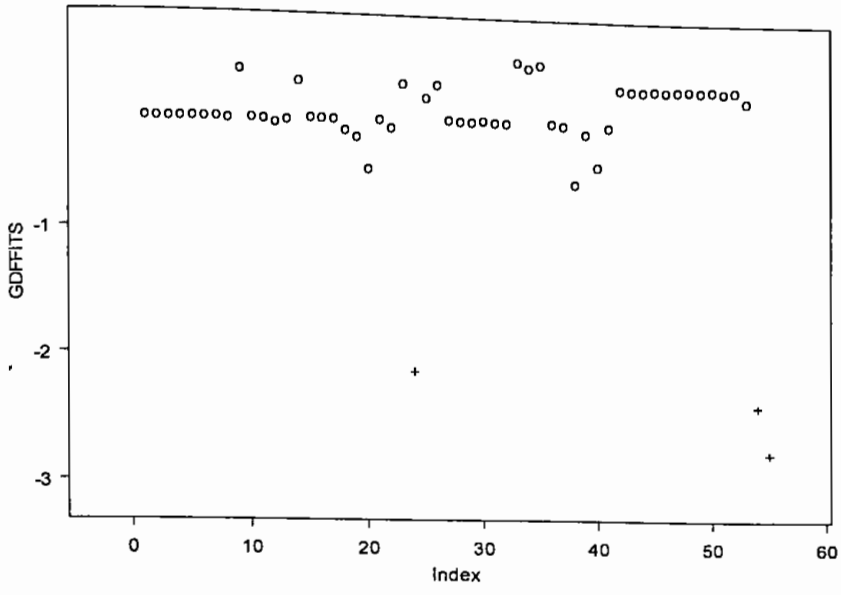
Modified Brown Data

The discussion about modified Brown data is given by the 2nd example in the section 6.2.4. Now we apply our two newly proposed group deletion measures *GDFFITS* and *GSDFBETA*. Both of the methods identify all of the three suspect cases as influential properly. Besides that

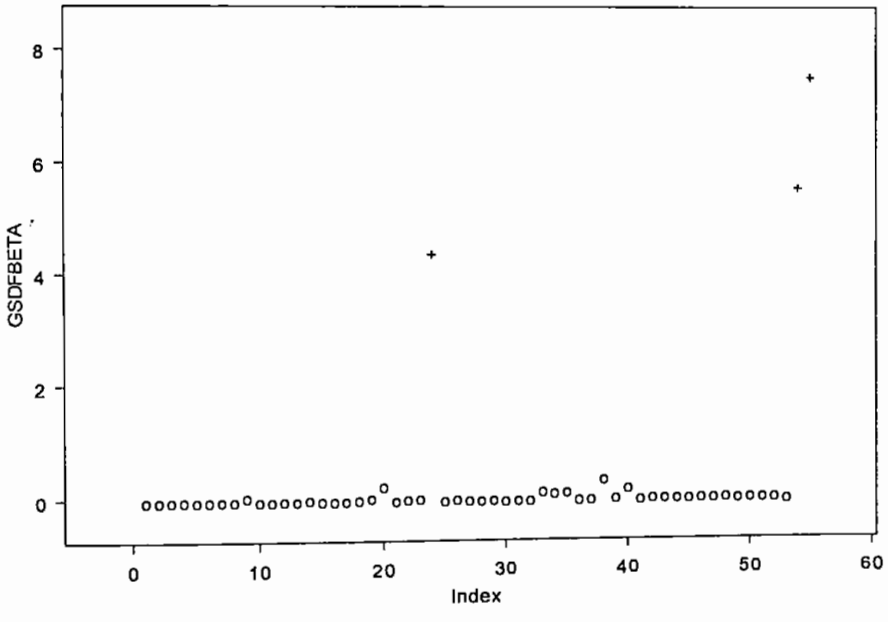
both identify one more case (38) as influential which makes the decision that the case (38) might be masked before the deletion of original 3 influential cases.

Table 6.3 Proposed diagnostic measures for modified Brown data

Index	|GDFFITS| (0.588)	|GSDFBETA| (0.367)	Index	|GDFFITS| (0.588)	|GSDFBETA| (0.367)
1	-0.0969	0.0097	29	-0.0984	0.0100
2	-0.1021	0.0107	30	-0.0892	0.0083
3	-0.0984	0.0100	31	-0.1015	0.0106
4	-0.0996	0.0102	32	-0.1045	0.0112
5	-0.0984	0.0100	33	0.3944	0.1613
6	-0.0977	0.0099	34	0.3498	0.1264
7	-0.0953	0.0094	35	0.3791	0.1488
8	-0.1082	0.0120	36	-0.0969	0.0097
9	0.2851	0.0836	37	-0.1098	0.0123
10	-0.1015	0.0106	38	**-0.5811**	**0.3710**
11	-0.1082	0.0120	39	-0.1711	0.0302
12	-0.1369	0.0192	40	-0.4386	0.2076
13	-0.1141	0.0133	41	-0.1167	0.0139
14	0.1993	0.0406	42	0.1932	0.0391
15	-0.0962	0.0096	43	0.1903	0.0377
16	-0.0977	0.0099	44	0.1864	0.0358
17	-0.0984	0.0100	45	0.1905	0.0372
18	-0.1892	0.0370	46	0.1876	0.0364
19	-0.2453	0.0629	47	0.1905	0.0372
20	-0.4976	0.2694	48	0.1993	0.0406
21	-0.0996	0.0102	49	0.1913	0.0381
22	-0.1630	0.0274	50	0.1993	0.0406
23	0.1912	0.0399	51	0.1874	0.0360
24	**-2.1214**	**4.3782**	52	0.1964	0.0410
25	0.0860	0.0082	53	0.1160	0.0150
26	0.1913	0.0381	**54**	**-2.3757**	**5.5408**
27	-0.0892	0.0083	**55**	**-2.7589**	**7.5386**
28	-0.0984	0.0100			



(a)



(b)

Figure 6.3(a) Index plot of GDFFITS, (b) Index plot of GSDFBETA

Modified Finney data

We now consider another data set given by Finney (1947). The original data set was obtained to study the effect of the rate and volume of air inspired on a transient vaso-constriction in the skin of the digits. The nature of the measurement process was such that only the occurrence and nonoccurrence of vaso-constriction could be reliably measured. We modify the data by putting five more outliers (cases 3, 4, 10, 11, 18, 20 and 21) where occurrence and nonoccurrence are replaced with each other. The modified data set is presented in table 5, appendix (A). Following scatter type character plot gives us the 4 dimensional information of the data set. The index of the cases are given by using respective numbers to the cases and colors red and blue respectively mean occurrence and nonoccurrence of vaso-constriction.

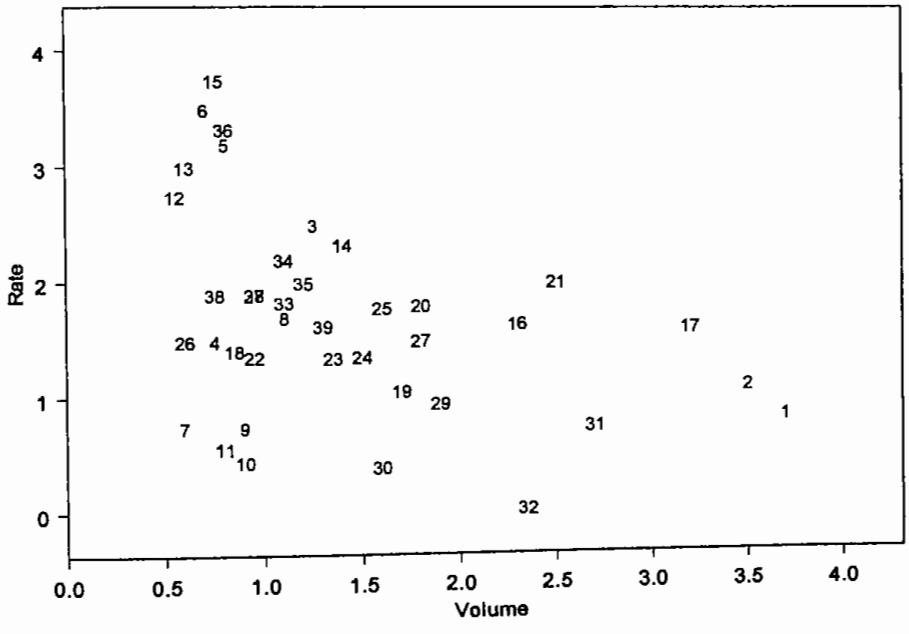


Figure 6.4 Character plot of modified Finney data

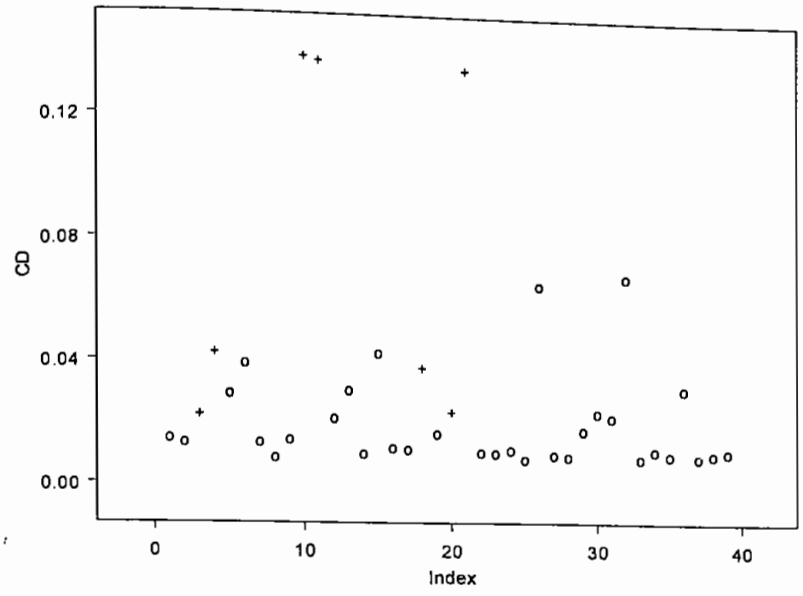
Application of single case deletion measures

This data set was analyzed extensively by Pregibon (1981). Looking at the pattern of occurrence and nonoccurrence with 25% and 75% contours in relation to rate and volume of the original data, Pregibon (1981) pointed out that this data set might contain two outliers (cases 4 and 18) in it.

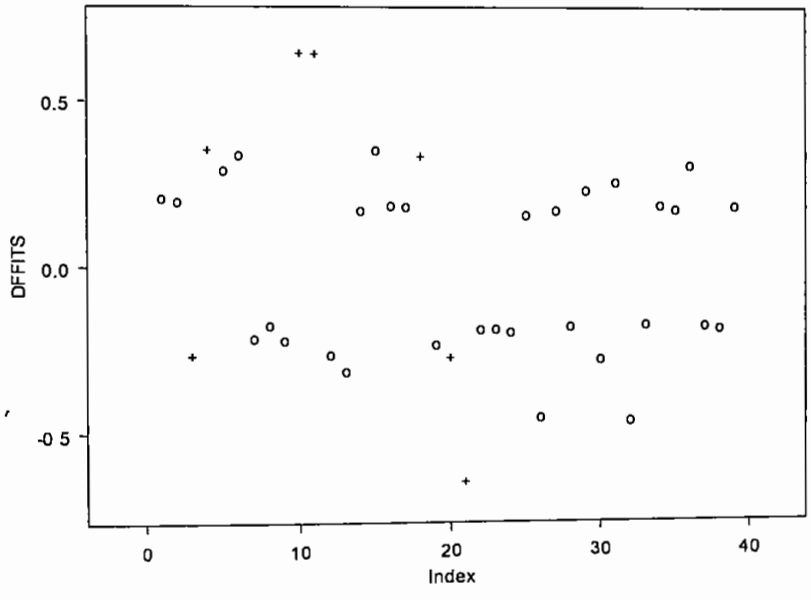
Table 6.4 presents influence diagnostics for the modified Finney data. We see from this table that the singe-deletion diagnostics *CD* and *DFFITS* fail totally to detect the influential observations.

Table 6.4 Influence diagnostics for modified Finney data

Index	CD (1.0)	DFFITS (0.832)	Ind.	CD (1.0)	DFFITS (0.832)
1	0.0154	0.2147	**21**	0.1362	-0.6392
2	0.0141	0.2060	22	0.0110	-0.1820
3	0.0228	-0.2613	23	0.0108	-0.1796
4	0.0431	0.3596	24	0.0119	-0.1886
5	0.0300	0.3001	25	0.0089	0.1634
6	0.0400	0.3464	26	0.0663	-0.4460
7	0.0143	-0.2068	27	0.0104	0.1767
8	0.0093	-0.1674	28	0.0099	-0.1726
9	0.0151	-0.2128	29	0.0186	0.2360
10	0.1398	0.6477	30	0.0245	-0.2711
11	0.1386	0.6449	31	0.0228	0.2615
12	0.0221	-0.2575	32	0.0695	-0.4567
13	0.0312	-0.3060	33	0.0094	-0.1683
14	0.0106	0.1781	34	0.0122	0.1913
15	0.0434	0.3609	35	0.0106	0.1785
16	0.0124	0.1927	36	0.0325	0.3122
17	0.0119	0.1890	37	0.0099	-0.1726
18	0.0384	0.3395	38	0.0110	-0.1813
19	0.0171	-0.2265	39	0.0116	0.1865
20	0.0239	-0.2680			



(a)



(b)

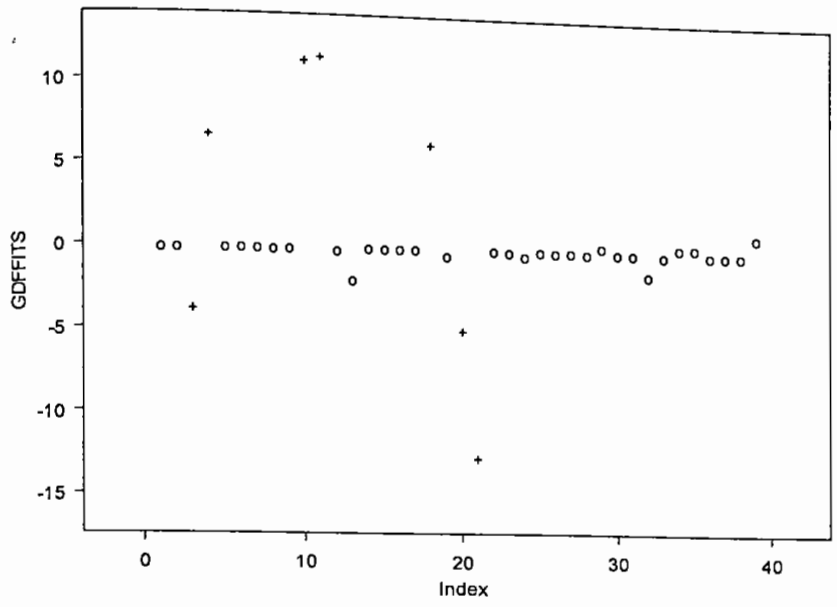
Figure 6.5 (a) Index plot of Cook's distance, (b) Index plot of DFFITS

Application of group deletion measures

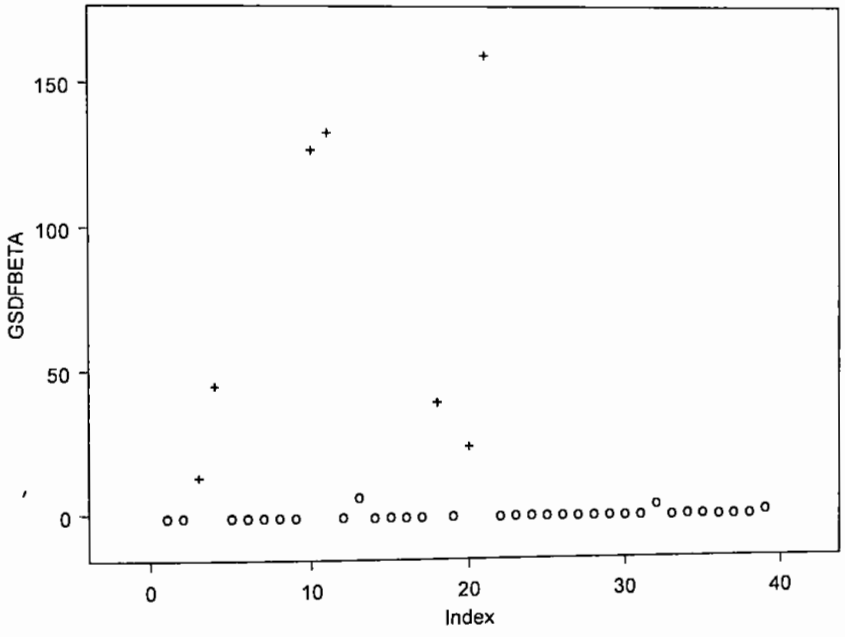
Our newly proposed *GDFFIT* and *GSDFBETA* successfully identify all 7 influential cases and at the same time both of them swamp 3 same cases (13, 32 and 39). Figures 6.6 (a and b) show the same evidence clearly.

Table 6.5 Influence diagnostics GDFFITS and GSDFBETA for modified Finney data

Index	GDFFITS (0.918)	GSDFBETA (1.038)	Index	GDFFITS (0.918)	GSDFBETA (1.038)
1	0.0000	0.0000	**21**	**-12.6585**	**160.2379**
2	0.0000	0.0000	22	-0.0001	0.0000
3	**-3.7293**	**13.6700**	23	-0.0467	0.0024
4	**6.7722**	**45.8560**	24	-0.2572	0.0777
5	0.0065	0.0000	25	0.0189	0.0004
6	0.0018	0.0000	26	0.0000	0.0000
7	0.0000	0.0000	27	0.0126	0.0002
8	-0.0293	0.0009	28	-0.0202	0.0004
9	0.0000	0.0000	29	0.3810	0.2080
10	**11.3037**	**127.7730**	30	-0.0003	0.0000
11	**11.5619**	**133.6776**	31	0.0000	0.0000
12	-0.1454	0.0270	**32**	**-1.2930**	**3.8080**
13	**-1.9011**	**6.8922**	33	-0.0886	0.0090
14	0.0023	0.0000	34	0.4066	0.2162
15	0.0001	0.0000	35	0.4288	0.2284
16	0.0000	0.0000	36	0.0019	0.0000
17	0.0000	0.0000	37	-0.0202	0.0004
18	**6.3102**	**39.8080**	38	-0.0010	0.0000
19	-0.3332	0.1349	**39**	**1.1413**	**1.5432**
20	**-4.9304**	**24.2208**			



(a)



(b)

Figure 6.6(a) Index plot of GDFFITS, (b) Index plot of GSDFBETA

Chapter 7

"A whole is what has beginning, middle and end."

Aristotle

Findings, Conclusions and Areas of Future Research

This chapter consists of three sections: the first and second ones depict our findings and conclusions, while the other presents areas of further research.

7.1 Findings

In this dissertation we endeavor to make three contributions: identification of influential observations in linear regression, classification of unusual observations in linear regression and identification of influential observations in logistic regression.

7.1.1 Identification of Influential Observations in Linear Regression

At the name of the topic, 'Identification of Influential Observations in Linear Regression', we develop two diagnostic measures: one is generalized SDFBETA (GSDFBETA), originated from squared difference in beta (SDFBETA), based on the idea of group deletion diagnostic, and the other is a new measure M_i that is developed by the idea of Pena (2005). We see that SDFBETA performs as well as a single case deletion measure like cook's distance and DFFITS and generalized measure GSDFBETA can successfully identify multiple influential observations even in presence of masking and/or swamping phenomena.

The new measure M_i extends the idea of Pena (2005) to the group deletion concept. It measures the influence of an observation based on how this observation is being influenced by the rest of the data. A number of well-referred data sets support the merit of our proposed method for the identification of influential observations where the commonly used methods fail. Moreover, this method is quite effective in finding influential observations from high dimensional large data sets and when there is evidence of variance heterogeneity in the data set. A simulation study shows the efficient performance of the new measure.

7.1.2 Classification of Unusual Observations in Linear Regression

Under this topic we have proposed a new type of plotting procedure that is exploratory in nature and which is able to identify and classify the multiple unusual observations: outliers, high-leverage points and influential observations at a time in a same graph. This method can make the five-fold plot by only one click in our computer key. Applications of our proposed method on a number of well-referred data sets show its efficient performance. Moreover, this method performs well for identifying unusual observations in case of high-dimensional large data set. We think this plotting technique may be a good addition to diagnostic literature and in statistical software packages for identification and at the same time for classification tasks of multiple unusual observations in linear regression.

7.1.3 Identification of Influential Observations in Logistic Regression

We also propose two new methods GDFFITS and GSDFBETA for the identification of influential observations in logistic regression, originated from the same measures of linear regression and based on the idea of group deletion. We see, most of the time the new measures identify the multiple influential observations properly even in presence misclassification of the response variable of the binomial logistic regression model, where as we show the existing methods almost fail to do so. Hence the methods are based

on group deletion of the suspect cases show nice performance in presence of masking and/or swamping phenomena. A number of examples have made the success story of the proposed methods.

7.2 Conclusions

This dissertation shows that proposed diagnostic measures successfully identify influential cases for linear and logistic regression in presence of masking and swamping. The methods based on robust regression in their construction process make them reliable in sense of robustness. Classification task for unusual observations in linear regression can give the accurate ideas about their consequences on the analysis and the decision making process. This study reaches the conclusion, 'robust regression and regression diagnostics are two complementary approaches and anyone is not good enough without the other'.

7.3 Areas of Future Research

I intend to continue the present work here along several main directions.

First of all, we intend to extend our proposed diagnostic measures to identify influential observations for high dimensional large data sets in data mining contexts. That may helps us to construct outliers free and fair observations in predictive modeling.

We plan to extend the identification techniques of influential observations for the Generalized Linear Model (GLM) (McCullagh and Nelder, 1989), multivariate, and non-linear regression.

We want to extend our ideas of classification of unusual observations in linear regression to the logistic regression and for GLM and we hope it also be applicable in non-linear regression.

We want to extend the view that identification of outliers or unusual observations in logistic regression can be performed for the classification tasks in data mining and pattern recognition. The tasks of classification by logistic regression may be extended to the area of kernel logistic regression (KLS), support vector machine (SVM) and import vector machine (IVM) for performing multi class and non-linear classification.

Appendix A

Data Sets Used in the Thesis

Table A.1 Monthly Payments Data

Month (x)	Payment (y)
1	3.22
2	9.62
3	4.50
4	4.94
5	4.02
6	4.20
7	11.24
8	4.53
9	3.05
10	3.76
11	4.23
12	42.69

Source: Rousseeuw et al., (1984a)

Table A.2 Hawkins *et al.* (1984) Data

Index	Y	X_1	X_2	X_3	Index	Y	X_1	X_2	X_3
1	9.7	10.1	19.6	28.3	39	-0.7	2.1	0	1.2
2	10.1	9.5	20.5	28.9	40	-0.5	0.5	2	1.2
3	10.3	10.7	20.2	31	41	-0.1	3.4	1.6	2.9
4	9.5	9.9	21.5	31.7	42	-0.7	0.3	1	2.7
5	10	10.3	21.1	31.1	43	0.6	0.1	3.3	0.9
6	10	10.8	20.4	29.2	44	-0.7	1.8	0.5	3.2
7	10.8	10.5	20.9	29.1	45	-0.5	1.9	0.1	0.6
8	10.3	9.9	19.6	28.8	46	-0.4	1.8	0.5	3
9	9.6	9.7	20.7	31	47	-0.9	3	0.1	0.8
10	9.9	9.3	19.7	30.3	48	0.1	3.1	1.6	3
11	-0.2	11	24	35	49	0.9	3.1	2.5	1.9
12	-0.4	12	23	37	50	-0.4	2.1	2.8	2.9
13	0.7	12	26	34	51	0.7	2.3	1.5	0.4
14	0.1	11	34	34	52	-0.5	3.3	0.6	1.2
15	-0.4	3.4	2.9	2.1	53	0.7	0.3	0.4	3.3
16	0.6	3.1	2.2	0.3	54	0.7	1.1	3	0.3
17	-0.2	0	1.6	0.2	55	0	0.5	2.4	0.9
18	0	2.3	1.6	2	56	0.1	1.8	3.2	0.9
19	0.1	0.8	2.9	1.6	57	0.7	1.8	0.7	0.7
20	0.4	3.1	3.4	2.2	58	-0.1	2.4	3.4	1.5
21	0.9	2.6	2.2	1.9	59	-0.3	1.6	2.1	3
22	0.3	0.4	3.2	1.9	60	-0.9	0.3	1.5	3.3
23	-0.8	2	2.3	0.8	61	-0.3	0.4	3.4	3
24	0.7	1.3	2.3	0.5	62	0.6	0.9	0.1	0.3
25	-0.3	1	0	0.4	63	-0.3	1.1	2.7	0.2
26	-0.8	0.9	3.3	2.5	64	-0.5	2.8	3	2.9
27	-0.7	3.3	2.5	2.9	65	0.6	2	0.7	2.7
28	0.3	1.8	0.8	2	66	-0.9	0.2	1.8	0.8
29	0.3	1.2	0.9	0.8	67	-0.7	1.6	2	1.2
30	-0.3	1.2	0.7	3.4	68	0.6	0.1	0	1.1
31	0	3.1	1.4	1	69	0.2	2	0.6	0.3
32	-0.4	0.5	2.4	0.3	70	0.7	1	2.2	2.9
33	-0.6	1.5	3.1	1.5	71	0.2	2.2	2.5	2.3
34	-0.7	0.4	0	0.7	72	-0.2	0.6	2	1.5
35	0.3	3.1	2.4	3	73	0.4	0.3	1.7	2.2
36	-1	1.1	2.2	2.7	74	-0.9	0	2.2	1.6
37	-0.6	0.1	3	2.6	75	0.2	0.3	0.4	2.6
38	0.9	1.5	1.2	0.2					

Source: Hawkins et al. (1984)

Table A.3 Stack loss Data

Index	Stack loss	Rate	Temperature	Acid Conc.
1	42	80	27	89
2	37	80	27	88
3	37	75	25	90
4	28	62	24	87
5	18	62	22	87
6	18	62	23	87
7	19	62	24	93
8	20	62	24	93
9	15	62	23	87
10	14	58	18	80
11	14	58	18	89
12	13	58	17	88
13	11	58	18	82
14	12	58	19	93
15	8	50	18	89
16	7	50	18	86
17	8	50	19	72
18	8	50	19	89
19	9	50	20	80
20	15	56	20	82
21	15	70	20	91

Source: Brownlee (1965)

Table A. 4 Hurtzs-Prung Russel Diagram Data

Index (i)	Log T. (X_i)	Log L. In. (Y_i)	Index (i)	Log T. (X_i)	Log L. In. (Y_i)
1	4.37	5.23	25	4.38	5.02
2	4.56	5.74	26	4.42	4.66
3	4.26	4.93	27	4.29	4.66
4	4.56	5.74	28	4.38	4.9
5	4.3	5.19	29	4.22	4.39
6	4.46	5.46	30	3.48	6.05
7	3.84	4.65	31	4.38	4.42
8	4.57	5.27	32	4.56	5.1
9	4.26	5.57	33	4.45	5.22
10	4.37	5.12	34	3.49	6.29
11	3.49	5.73	35	4.23	4.34
12	4.43	5.45	36	4.62	5.62
13	4.48	5.42	37	4.53	5.1
14	4.01	4.05	38	4.45	5.22
15	4.29	4.26	39	4.53	5.18
16	4.42	4.58	40	4.43	5.57
17	4.23	3.94	41	4.38	4.62
18	4.42	4.18	42	4.45	5.06
19	4.23	4.18	43	4.5	5.34
20	3.49	5.89	44	4.45	5.34
21	4.29	4.38	45	4.55	5.54
22	4.29	4.22	46	4.45	4.98
23	4.42	4.42	47	4.42	4.5
24	4.49	4.85			

Source: Rousseeuw and Leroy (1987)

Table A. 5 Aircraft Data

Index	Cost	Aspect Ratio	Lift-to-drag Ratio	Weight	Thrust
1	2.76	6.3	1.7	8176	4500
2	4.76	6.0	1.9	6699	3120
3	8.75	5.9	1.5	9663	6300
4	7.78	3.0	1.2	12837	9800
5	6.18	5.0	1.8	10205	4900
6	9.5	6.3	2.0	14890	6500
7	5.14	5.6	1.6	13836	8920
8	4.76	3.6	1.2	11628	14500
9	16.7	2.0	1.4	15225	14800
10	27.68	2.9	2.3	18691	10900
11	26.64	2.2	1.9	19350	16000
12	13.71	3.9	2.6	20638	16000
13	12.31	4.5	2.0	12843	7800
14	15.73	4.3	9.7	13384	17900
15	13.59	4.0	2.9	13307	10500
16	51.9	3.2	4.3	29855	24500
17	20.78	4.3	4.3	29277	30000
18	29.82	2.4	2.6	24651	24500
19	32.78	2.8	3.7	28539	34000
20	10.12	3.9	3.3	8085	8160
21	27.84	2.8	3.9	30328	35800
22	107.1	1.6	4.1	46172	37000
23	11.19	3.4	2.5	17836	19600

Source: Gray (1985)

Table A. 6 Brown Data

Index	L.N.I	A.P.	Index	L.N.I	A.P.
1	0	48	28	0	50
2	0	56	29	0	50
3	0	50	30	0	40
4	0	52	31	0	55
5	0	50	32	0	59
6	0	49	33	1	48
7	0	46	34	1	51
8	0	62	35	1	49
9	1	56	36	0	48
10	0	55	37	0	63
11	0	62	38	0	102
12	0	71	39	0	76
13	0	65	40	0	95
14	1	67	41	0	66
15	0	47	42	1	84
16	0	49	43	1	81
17	0	50	44	1	76
18	0	78	45	1	70
19	0	83	46	1	78
20	0	98	47	1	70
21	0	52	48	1	67
22	0	75	49	1	82
23	1	99	50	1	67
24	0	187	51	1	72
25	1	136	52	1	89
26	1	82	53	1	126
27	0	40			

Source: Brown et al. (1980)

Table A. 7 Modified Brown Data

Index	L.N.I.	A.P.	Index	L.N.I.	A.P.
1	0	48	29	0	50
2	0	56	30	0	40
3	0	50	31	0	55
4	0	52	32	0	59
5	0	50	33	1	48
6	0	49	34	1	51
7	0	46	35	1	49
8	0	62	36	0	48
9	1	56	37	0	63
10	0	55	38	0	102
11	0	62	39	0	76
12	0	71	40	0	95
13	0	65	41	0	66
14	1	67	42	1	84
15	0	47	43	1	81
16	0	49	44	1	76
17	0	50	45	1	70
18	0	78	46	1	78
19	0	83	47	1	70
20	0	98	48	1	67
21	0	52	49	1	82
22	0	75	50	1	67
23	1	99	51	1	72
24	0	187	52	1	89
25	1	136	53	1	126
26	1	82	54	0	200
27	0	40	55	0	220
28	0	50			

Source: Brown et al. (1980)

Table A. 8 Modified Finney Data

Index	Response	Volume	Rate	Index	Response	Volume	Rate
1	1	3.70	0.820	21	0	2.50	2.000
2	1	3.50	1.090	22	0	0.95	1.360
3	0	1.25	2.500	23	0	1.35	1.350
4	1	0.75	1.500	24	0	1.50	1.360
5	1	0.80	3.200	25	1	1.60	1.780
6	1	0.70	3.500	26	0	0.60	1.500
7	0	0.60	0.750	27	1	1.80	1.500
8	0	1.10	1.700	28	0	0.95	1.900
9	0	0.90	0.750	29	1	1.90	0.950
10	1	0.90	0.450	30	0	1.60	0.400
11	1	0.80	0.570	31	1	2.70	0.750
12	0	0.55	2.750	32	0	2.35	0.030
13	0	0.60	3.000	33	0	1.10	1.830
14	1	1.40	2.330	34	1	1.10	2.200
15	1	0.75	3.750	35	1	1.20	2.000
16	1	2.30	1.640	36	1	0.80	3.330
17	1	3.20	1.600	37	0	0.95	1.900
18	1	0.85	1.420	38	0	0.75	1.900
19	0	1.70	1.060	39	1	1.30	1.630
20	0	1.80	1.800				

Source: Finney (1947)

Appendix B

Abstracts of Published, Accepted and Submitted Articles

(Related to the Dissertation)

Application of Robust Regression and Bootstrapping in Purchasing Power Parity Analysis

A. A. M. Nurunnabi*, Mohammed Nasser**

Abstract: This article is an attempt to show how robust regression, a computer based statistical technique introduced by P.J.Huber in 1973 and later developed by Rousseeuw (1984), Rousseeuw and Yohai (1984), and many others, can helps us in cases where OLS totally fails due to outliers, leverage points and non-normality of error distribution. But to infer from the estimators obtained from robust regression we generally need, especially for small samples, bootstrapping (resampling) technique that is also a computer intensive statistical technique introduced by Efron (1979), and later developed in many directions. This talk illustrates the whole thing by an example using data extracted from the Big Mac. Index, with a purchasing power parity analysis.

*Assistant Professor, Department of Business Administration, Uttara University, Dhaka-1230. ** Professor, Department of Statistics, University of Rajshahi, Rajshahi-6205

Article Published in
Daffodil Intl. University Journal of
Business and Economics,
Vol. 2, No. 1, January 2007

Knowledge Inequality between Male and Female on HIV/AIDS in Bangladesh

A.A.M. Nurunnabi[1], A.H.M. Rahmatullah Imon[2], and Mohammed Nasser[2]

[1]Assistant Professor, Department of Business Administration,
Uttara University, Dhaka – 1230,
[2]Professor, Institute for Mathematical Research,
University Putra Malaysia, 43400 Serdang, Selangor, Malaysia,
[3] Professor, Department of Statistics, University of Rajshahi, Rajshahi-6205

ABSTRACT

Globally women are becoming infected with HIV at a faster rate than men. Women accounted for nearly 41% of all people living with HIV worldwide in 1997, but this figure increased up to more than 50% by 2004. It is generally believed that the challenges to fight against HIV/AIDS are closely related with many economic and social factors of a country. But it is now evident that lack of knowledge and awareness about the causes and preventions regarding HIV/AIDS can magnify the risk of infection to a greater extent. Bangladesh is one of the least developed countries with a very low literacy rate and it has gender inequality in almost every respect. In this paper we tried to show that the knowledge as well as awareness of HIV/AIDS differs significantly between women and men based on the data extracted from Bangladesh Demographic and Health Survey report 2004.

Keywords: AIDS; BDHS; Exploratory data analysis; HIV; Knowledge difference; Logistic regression.

Article accepted in South Asian Journal of Population and Health Vol. 1, No. 1, January 2008

APPLICATIONS OF ROBUST REGRESSION IN BUSINESS, ECONOMICS AND SOCIAL SCIENCES

A.A. M. Nurunnabi
Department of Business Administration, Uttara University, Dhaka-1230
Mohammed Nasser
Department of Statistics, University of Rajshahi, Rajshahi-6205
A.H.M. Rahmatullah Imon
Institute of Mathematical Research, University Putra Malaysia,
43400 Serdang, Selangor, Malaysia

Abstract

Robust regression techniques are rarely used in business, economics or in social sciences. It is a reliable alternative, where ordinary least squares (OLS) totally fails due to unusual observations and the violations of normality assumptions of error distributions. We demonstrate the importance of robust regression techniques by studying and comparing with OLS. Three examples are taken from the literature in areas of business, economics and social sciences.

Keywords: Influential observation; Least median of squares; Least trimmed squares; Leverage point; Outlier; Reweighted least squares.

Article Accepted for
The Journal of NSU School of Business,
North South Business Review, Vol. 1, No. 2, 2007

A New Measure for the Identification of Influential Observations in Linear Regression

A.A.M. Nurunnabi[1], A.H.M. Rahmatullah Imon[2], Mohammed Nasser[3]

[1]Department of Business Administration, Uttara University, Dhaka-1230, Bangladesh

[2,3]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh

Abstract

In linear regression it is a common practice of measuring influence of an observation is to delete the case from the analysis and to investigate the norm of the change in the parameters or in the vector of forecasts resulting from this deletion. Pena (2005) introduced a new idea to measure the influence of an observation based on how this observation is being influenced by the rest of the data. In this article we would like to extend this idea to a group deletion technique suggested by Hadi and Simonoff (1993) and propose a new statistic to identify influential observations in linear regression. We investigate the usefulness of the proposed technique by two well-referred data sets and an artificial data with high-dimension, heterogeneous variances and large number of observations.

Key Words: Influential observations; group-deleted measure; masking; swamping; high dimensional large data; heterogeneous variances.

Article Presented at the 10th National Statistical Conference; Dhaka 2007

Classification of Outliers, High-Leverage Points and Influential Observations in Linear Regression

By A. A. M. NURUNNABI

Uttara University, Dhaka-1230, Bangladesh

M. NASSER and A.H.M. R. IMON

University of Rajshahi, Rajshahi-6205, Bangladesh

SUMMARY

In this paper we propose a five-fold plotting technique with a robust distance measure on a potential-residual (P-R) plot that can identify and classify outliers, high leverage points and influential observations at the same time in a same graph. The proposed technique based on group deletion idea shows efficient performance in presence of masking and/or swamping phenomena. We demonstrate the proposed technique by using three well-referred data sets and an artificial high-dimensional large data with heterogeneous variances.

Keywords: INFLUENTIAL OBSERVATION; LEVERAGE POINT; MAHALANOBIS DISTANCE; MASKING; OUTLIER; POTENTIAL-RESIDUAL PLOT; SWAMPING

Article Submitted to the Journal of the Royal Statistical Society

Identification of Multiple Influential Observations in Logistic Regression

A.A. M. Nurunnabi[a,*], **A. H. M. Rahmatullah Imon**[b], **Mohammed Nasser**[b]
[a] *Department of Business Administration, Uttara University, Dhaka – 1230, Bangladesh*
[b] *Department of Statistics, University of Rajshahi, Rajshahi -6205, Bangladesh*

ABSTRACT *The identification of influential observations in logistic regression has drawn a great deal of attention in recent years. Most of the available techniques like Cook's distance and DFFITS are based on single case deletion. But there is evidence that these techniques suffer from masking and swamping problems and consequently fail to detect multiple influential observations. In this paper an attempt has been made to develop a new measure for the identification of multiple influential observations based on a generalized version of DFFITS. The advantage of using the proposed method in the identification of multiple influential cases is then investigated through several well-referred data sets.*

KEY WORDS: Influential observation, High leverage point, Outlier, Masking, Swamping, Generalized DFFITS

Article Prepared for Submission
In the Journal of Applied Statistics

Bibliography

Agullo, J. (1996) Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm, *Proceedings in Computational Statistics,* Ed. Prat. A.Heidelberg: Physica-Verlag, 175-180.

Allen, D. M. (1974) The Relationship between variable selection and data augmentation and a method of prediction, *Technometrics,* 16, 125-127.

Andrews, D.F. and Pregibon, D. (1978) Finding the outliers that matter, *Journal of the Royal Statistical Society,* B, 40, 85-93.

Atkinson, A.C. (1981) Two graphical displays for outlying and influential observations in regression, *Biometrika*, 68, 13 – 20.

Atkinson, A. C. (1982) Regression diagnostics, transformations and constructed variables, *Journal of the Royal Statistical Society*, B, 44, 1-36.

Atkinson, A.C. (1985) *Plots, Transformations and Regression*, Oxford: Clarendon Press.

Atkinson, A.C. (1986) Masking unmasked, *Biometrika*, 73, 533-541.

Atkinson, A.C., and Mulira, H. M. (1993) The stalactite plot for the detection of multivariate outliers, *Statistics and Computing,* 3, 27-35.

Atkinson, A.C. (1994) Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, 89, 1329 – 1339.

Atkinson, A.C. and Riani, M. (2000) *Robust Diagnostic Regression Analysis*, New York: Springer.

Barnett, V. and Lewis, T.B. (1995) *Outliers in Statistical Data,* New York: Wiley.

Basilevsky, A. (1983) *Applied Matrix Algebra in the Statistical Science*, New York: North Holland.

Becker, C. and Gather, U. (1999) The masking breakdown point of multivariate outlier identification rules, *Journal of the American Statistical Association,* 94, 947-955.

Beckman, R. J., and Trussell, H. J. (1974) The distribution of the arbitrary studentized residual and the effects of updating in multiple regression, *Journal of the American Statistical Association,* 69, 199-201.

Behnken, D.W. and Draper, N.R. (1972) Residuals and their variance patterns, *Technometrics,* 14, 101-111.

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

Bingham, C. (1977) Some identities useful in the analysis of residuals from linear regression, *Technical Report* 300, School of Statistics, University of Minnesota.

Brown, B.W., Jr. (1980) Prediction analysis for binary data, in *Biostatistics Casebook*, R.G. Miller, Jr., B. Efron, B. W. Brown, Jr., L.E. Moses (Eds.), New York: Wiley.

Brownlee, K. A. (1965) *Statistical Theory and Methodology in Science and Engineering,* New York: Wiley.

Box, G. E. P. (1953) Non-normality and tests on variances, *Biometrics*, 40, 318-335.

Butler, R. W., Davies, P. L. and Jhun, M. (1990) Asymptotics for the minimal covariance determinant estimator, *The Annals of Statistics*, 21, 1385-1400.

Chatterjee, S. and Hadi, A.S. (1986) Influential observations, high leverage points, and outliers in regression, *Statistical Sciences*, 1, 379-416.

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

Chatterjee, S. and Hadi, A.S. (2006) *Regression Analysis by Examples*, New York: Wiley.

Chambers, J.M. and Hastie, T.J. (1992) *Statistical Models in S*, Wadsworth and Brooks, Pacific Grove, CA.

Cook, R.D. (1977) Detection of influential observations in linear regression, *Technometrics*, 19, 15-18.

Cook, R.D. (1979) Influential observations in regression, *Journal of the American Statistical Association*, 74, 169-174.

Cook, R. D. (1998) *Regression Graphics; Ideas for Studying Regression Through Graphics,* New York: Wiley.

Cook, R.D., and Weisberg, S. (1980) Characterization of an empirical influence functions for detecting influential cases in regression, *Technometrics,* 22, 495-508.

Cook, R.D., and Weisberg, S. (1982) *Residuals and Influence in Regression*, London: Chapman and Hall.

Cookley, C. W., and Hettmans perger, T. P. (1993) A bounded influence, high breakdown, efficient regression estimator, *Journal of the American Statistical Association,* 88, 872-880.

Daniel, C. and Wood, F. S. (1971) *Fitting Equations to Data,* New York: Wiley.

Davis, R. B. and B. Hutton (1975) The effects of errors in the independent variables in linear regression, *Biometrika*, 62, 383-391.

Davies, L., and Gather, U. (2004) Robust Statistics, Examples and Introduction, http://www.quantlet.com/indstat/scripts/csa/html/model175/.html, accessed on 22 July 2007.

Dhrymes, P. J. (1984) *Mathematics for Econometrics*, New York: Springer Verlag.

Donoho, D. L. (1982) Breakdown properties of multivariate location estimators, *Qualifying paper,* Harvard University.

Donoho, D. L. and Huber, P. J. (1983) The notion of breakdown point. in: Bickel, P. J. and Hodges Jr., J.I. (Eds), *A Festschrift for L. Lehmann*, Wadsworth, Belmont, CA, 157-184.

Draper, N.R. and John, J.A. (1981) Influential observations and outliers in regression, *Technometrics*, 23, 21-26.

Dutter, R. (1977) Numerical solution of robust regression problems: Computaional aspects, a comparison, *Journal of Statistical Computation and Simulation*, 5, 207-238.

Ellenberg, J. H. (1976) Testing for a single outlier from a general regression, *Biometrics*, 32, 637-645.

Ellis, S. P., and Morgenthaler, S. (1992) Leverage and breakdown in L_1 regression, *Journal of the American Statistical Association*, 87, 143-148.

Efron, B. (1979) A Bootstrap methods: another look at the jackknife, *The Annals of statistics*, 7, 1-26.

Efron, B, and Gong, G. (1983) A leisurely look at the bootstrap, the jackknife and cross validation, *American Statistician*, 37, 36-*48*.

Efron, B., and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*, New York: Wiley.

Edgeworth, F. Y. (1887) On observations relating to several quantities, *Hermathena*, 6, 279-285.

Finney, D. J. (1947) The estimation from individual records of the relationship between dose and quantal response, *Biometrika*, 34, 320-334.

Fox, J. (1993) Regression diagnostics, in *Regression Analysis*, (editor, Lewis Beck, M. S.), UK: Sage Publications.

Gnanadesikan, R. and Kettenring, J. (1972) Robust estimates, residuals, and outlier detection with multi-response data, *Biometrics*, 28, 81-124.

Gray, G. B. (1985) Graphics for regression diagnostics, *American Statistical Association Proceedings of the Statistical Computing section*, 102-107.

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

Hadi, A. S. (1994) A modification of a method for the detection of outliers in multivariate samples, *Journal of the Royal Statistical Society,* B, 56.

Hadi, A.S. and Simonoff, J.S. (1993) Procedure for the identification of outliers in linear models, *Journal of the American Statistical Association*, 88, 1264 – 1272.

Hampel, F. R. (1968) Contributions to the theory of robust estimation, *Ph D Thesis*, University of California, Barkeley.

Hampel, F. R. (1975) Beyond location parameters: robust concepts and methods, Bull., Inst., 46, 375-382.

Hampel, F. R. (1974) The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, 62, 1179-1186.

Hampel, F. R., Marazzi, A. Ronchetti, E., Rousseeuw, P. J., Stahel, W. A., and Welsch, R. E. (1982) Handouts for the instructional meeting on robust statistical methods, 15th *European Meeting of Statisticians*, Palermo, Italy.

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

Hardin, J. and Rocke, D. M. (2005) The distribution of robust distances. *Journal of Computational and Graphical Statistics,* 14, 4, 928-946.

Hawkins, D. M. (1980) *Identification of Outliers*, London: Chapman Hall.

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.

He, X., Simpson, D. L., and Portnoy, S. L. (1990) Breakdown robustness of tests, *Journal of the American Statistical Association,* 85, 446-452.

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

Hinkly, D. V. (1977) On Jackknifing in unbalanced situations, *Technometrics*, 19, 285-292.

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

Hocking, R. R. (1983) Developments in Linear Regression Methodology: 1959-1982, *Technometrics,* 25, 219-249.

Hocking, R.R., and Pendleton, O.J. (1983) The regression dilemma, *Communications in Statistics- Theory and Methods*, 12, 497-527.

Holland, P. W., and Welsch. R. E. (1977) Robust regression using iteratively reweighted least-squares, *Communication in Statistics*, A 6, 813-888.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression*, New York: Wiley.

Hossjer, O. (1994) Rank-based estimates in the linear model with high breakdown point, *Journal of the American Statistical Association*, 89, 149-158.

Huber, P. J. (1964) Robust estimation of a location parameter, *Annals of Mathematical Statistics*, 35, 73-101.

Huber, P. J. (1973) Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, 1, 799-821.

Huber, P. J. (1975) *Robustness and Designs, A survey of Statistical Design and Linear Models*, Amsterdam: North-Holland.

Huber, P. J. (1977) Robust covariances, in statistical decision theory and related topics. Vol. 2., edited by S. S. Gupta and D. S. Moore, Academic Press. New York, 165-191.

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

Huber, P. J. (1991) Between robustness and diagnostics, in *Direction in Robust Statistics and Diagnostics* Part-I, Stahel, W. and Weisberg, S, editors, New York: Springer-Verlag, 121-130.

Hubert, M. (1997) The breakdown value of the L_1 estimator in contingency tables, *Statistics and Probability Letters*, 33, 419-425.

Hubert, M., Rousseeuw, P. J. and Aelst, S. V. (2007) High breakdown robust multivariate methods, *Statistical Science*, to appear.

Huber, P. J. and Dutter, R. (1974) Numerical solution of robust regression problems, COMPSTAT 1974, *Proceeding in Computational Statistics*, G. Bruckmann, F. Ferschl, and L. Schmetterer (ed.) Physica, Vienna, 165-172.

Imon, A. H. M. R. (1996) Sub-sample methods in regression residual prediction and diagnostics, Unpublished *Ph D thesis*, University of Birmingham, UK.

Imon, A. H. M. R. (2002) Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies,* Special Volume, 32, 73 – 90.

Imon, A. H. M. R. (2005) Identifying multiple influential observations in linear regression, *Journal of Applied Statistics*, 32, 73 – 90.

Imon, A. H. M. R. (2006) Identification of high leverage points in logistic regression, *Pakistan Journal of Statistics*, 22, 147 – 156.

Imon, A.H.M.R. and Hadi, A.S. (2005) Identification of multiple outliers in logistic regression, Paper presented in the *International Statistics Seminar on 'Statistics in the Technological Age'*, Institute of Mathematical Statistics, University of Malaya, KualaLumpur.

Imon, A.H.M.R. and Ali, M. M. (2005) Simultaneous identification of multiple outliers and high leverage points in linear regression, *Journal of Korean Data and Information Science Society,* 16, 2, 429-444.

Mahalanobis, P. C. (1936) On the generalized distance in statistics, *Proceedings of the National Institute of Science of India,* 12, 49-55.

Mallows, C. L. (1975) On Topics in Robustness, *Technical Memorandum*, Bell Telephone Laboratories, Murry Hill, N. J.

Mardia, K., Kent, J. and Bibby, J. (1979) *Multivariate Analysis*, New York: Academic Press.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006) *Robust Statistics*, New York: Wiley.

Maronna, R. A., Bustos, O., and Yohai, V. (1979), Bias and efficiency robustness of general M-estimators for regression with random carriers, in *Smoothing Techniques for Curve Estimation,* edited by T. Gadder and M. Rosenblatt, , New York: Springer Verlag, 91-116.

Mc. Cullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, London: Chapman and Hall.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001) *Introduction to Linear Regression Analysis*, New York: Wiley.

Mosteller, F., and Tukey, J. W. (1977) *Data Analysis and Regression*, Addison-Wesley, Reading, MA.

Marazzi, A. (1980), ROBETH, A subroutine library for robust statistical procedures, in COMPSTAT 1980, *Proceedings in Computational Statistics*, Physica Verlag, Vienna.

Miller, R. G. (1974) The Jackknife: a review, *Biometrika*, 61, 1-15.

Moore, E. H. (1920) On the reciprocal of the general algebraic matrix, Bulletin; *American Mathematical Society*, 26, 394-395.

Nasser, M. (2000) Continuity and differentiability of statistical functionals: its relation to robustness in bootstrapping, Un published *PhD Thesis*, RCMPS, University of Chittagong, Bangladesh.

Pena, D. (2005) A new statistic for influence in linear regression, *Technometrics*, 47, 1 – 12.

Pena, 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,* B, 57, 145-156.

Penrose, R. A. (1955) A generalized inverse for matrices, Proceeding; *Cambridge Philosophical Society*, 51, 406-413.

Pregibon, D. (1981) Logistic regression diagnostics, *Annals of Statistics*, 9, 977 – 986.

Quenouille, M. (1949) Approximate tests of correlation in time series, *Journal of the Royal Statistical Society*, B, 11, 18-44.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, New York: Wiley.

Rousseeuw, P.J. (1984) Least median of squares regression, *Journal of the American Statistical Association*, 79, 871 – 880.

Rousseeuw, P.J. (1985) A regression diagnostic for multiple outliers and leverage points, *Abstracts in IMS Bulletin*, 14, 399.

Rousseeuw, P. J., Daniel, B., and Leroy, A. (1984a) Applying robust regression to insurance, *Insur. Math. Econ.* 3, 67-72.

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

Rousseeuw, P. J., and Hubert, M. (1997) Recent developments in PROGRESS, in L_1 Statistical Procedures and Related Topics, (edited by Dodge, Y., *The IMS Lecture Notes* Monograph Series, Vol. 31, 201-214.)

Rousseeuw, P. J., and Yohai, V. (1984) Robust regression by means of S-estimators, in *Robust and Non-linear Time series Analysis,* (edited by Frank, W., Hardle, and Martin, R. D.) New York: Springer Verlag.

Rousseeuw, P. J. and van Zomeren, B. C. (1990) Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association,* 85, 633-639.

Rousseeuw, P. J. and van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41, 212-223.

Ryan, T.P. (1997) *Modern Regression Methods*, New York: Wiley.

Serfling, R. (1980) *Approximation Theorems of Mathematical Statistics*, New York: Wiley..

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*, New York: Springer-Verlag.

Simonoff, J. S. (1991) General approaches to stepwise identification of unusual values in data analysis, in *Robust Statistics and Diagnostics*: Part II, eds. W. Stahel and S. Weisberg, New York: Springer Verlag, 223-242.

Simpson, D. G., Ruppert, D., and Carroll, R. J. (1992) On one-step GM-estimates and stability of inference in linear regression, *Journal of the American Statistical Association*, 87, 439-450.

Stahel, W. A. (1981) Robust estimation: infinitesimal optimality and covariance matrix estimator, *PhD thesis*, German.

Stahel, W. and Weisberg, S. (1991) *Direction in Robust Statistics and diagnostics*, Part I & Part II, New York: Springer Verlag.

Staudte, R. G., and Sheather, S. J. (1990) *Robust Estimation and Testing*, New York: Wiley.

Tukey, J. W. (1958) Bias and Confidence in Not Quite Large Samples, (Abstract), *The Annals of Mathematical Statistics, 29, 614.*

Tukey, J. W. (1960) A survey of sampling from contaminated distributions, in Olkin, I (Ed.), *Contributions to probability and Statistics*, Stanford University Press, Stanford, California.

Tukey, J. W. (1962) The future of data analysis, *Annals of Mathematical Statistics,* 33, 1-67.

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

Welsch, R.E. (1982) Influence functions and regression diagnostics, in *Modern Data Analysis*, Launer, R.L. and Siegel, A.F. (Ed.), 149-169.

Welsch, R. E. and Kuh, E. (1977) *Linear Regression Diagnostics*, Sloan School of Management

Welsch, R. E. S. C. Peters (1978) Finding influential subsets of data in regression models, in A. R. Gallant and T. M. Gerig (Eds.), *Proceedings of the Eleventh Interface Symposium on computer Science and Statistics*, Institute of Statistics, North Carolina State University, 240-244.

Wilks, S. (1962) *Mathematical Statistics*, New York: Wiley.

Willems, G., and Aelst, S. V. (2004) Fast and robust bootstrap for LTS, in preprint, submitted to *Elsevier Science*.

Wu, C. F. J. (1986) Jackknife, bootstrap, and other resampling methods in regression analysis, *The Annals of Statistics*, 14, 1261-1350.

Yohai, V. J. (1985) High breakdown-point and high efficiency robust estimates for regression, to appear *The Annals of Statistics*.

Yohai, V. J. (1987) High breakdown point and high efficiency robust estimates for regression, *The Annals of Statistics*, 15, 642-656.