CLUSTER-BASED ESTIMATORS FOR MULTIPLE AND MULTIVARIATE LINEAR REGRESSION MODELS

EKELE ALIH

UNIVERSITI SAINS MALAYSIA

2015

CLUSTER-BASED ESTIMATORS FOR MULTIPLE AND MULTIVARIATE LINEAR REGRESSION MODELS

by

EKELE ALIH

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

June 2015

DEDICATIONS

Anabel-Martha Ugbedeojo, Mummy's death has made you unstable– I just wish I can make up for you upon my return

Gabriel-AkoEkele, Daddy left you to study just few hours after you came–I just hope you will recognize me when I return

. . .

ACKNOWLEDGEMENTS

I give honour, thanks, gratitude and appreciation to the Creator God for bringing me into this world, for being with me in times of trouble, for providing all my needs and above all for giving me sound wisdom, knowledge and understanding even in the face of uncertainties in the course of this Doctorate Degree Programme.

My special appreciation goes to Associate Professor Hong Choon Ong, who welcomed me like a father will welcome a son when I arrived in Malaysia; who guided me like a father will guide his son during my registration process to be a PhD student of USM; who provided for me, all I needed to do in my research when I eventually settled down as a PhD student of PPSM, USM; who took all the pains to go through every line of this Thesis and offered suggestions and corrections. I sincerely thank him for condoning all my acts, both '*the good*', '*the bad*' and '*the ugly*' and for heartily moderating this research project to its peak. Once, again, Remain blessed, Prof. Ong.

I thank my Parents, Mr. Stephen Ibrahim Alih and Mrs. Martha Ibrahim Alih, through whom I came into this world. They have persistently been praying for my success and hope to celebrate my victory in this PhD journey. Thank you, 'MUM&DAD'.

My wife made sacrifices in no small measure in the course of my PhD pursuit. One of those sacrifices was to allow me to travel for the PhD program barely 8 days after you underwent a caesarian operation that gave birth to our boy, Gab. You borne the pains of the wound and the pains of keeping the family together alone in my absence. This

is love at its peak. These gestures are indelible in my mind. Thank you, Stella-Gbelei.

The Federal Polytechnic Idah is immensely appreciated for providing my funding through the TETFUND of the Federal Republic of Nigeria. Thank you to the Government of '*myfatherland*'

I appreciate my two kids, Anabel and Gabriel, who missed me dearly and whom I dearly missed while I am away to pursue the PhD. I love you and will soon join you as we hope to grow together in God's path.

Colleagues and staff of Maths/Statistics Department, Federal Polytechnic Idah are also acknowledged for their co-operation with me while pursuing the programme. Thank you all.

Above all, I thank once more the Almighty God for His unquantifiable journey mercies and for sustaining my life all through the period of study. Once again, I acknowledge that I am divinely favoured.

Ekele Alih

2015

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3SLS 3 Stage Least Squares

AHC Agglomerative Hierarchical Clustering

ANOVA Analysis-of-Variance

BACON Block Adaptive Computationally-efficient Outlier Nominator

BDP Break Down Point

BI Bounded Influence

BIR Bounded Influence Regression

BLUE Best Linear Unbiased Estimator

CBI Cluster-Based Bounded Influence Regression

C-Step Concentration Steps

CL2RR Cluster-based L2 Robust Regression

CLreg Cluster-based Estimator for Multiple and Multivariate Linear Regression

CMRR Cluster-based Multivariate Robust Regression

CRLB Cramer-Rao Lower Bound

det Determinant of a Matrix

DetMCD Deterministic Minimum Covariance Determinant

df Degree of Freedom

DFFITS Difference in Fits influence diagnostics

DFFITS$_i$ Difference in Fits influence diagnostics when i^{th} observation has been removed

DFFITS$_l$ Difference in Fits influence diagnostics when the set of l observations has been removed

DHC Divisive Hierarchical Clustering

FastMCD Fast Minimum Covariance Determinant

FSA Forward Search Algorithm

GM Generalized Maximum Likelihood Estimator

iid Independently and identically distributed

IRLS Iteratively Reweighted Least Squares

LAD Least Absolute Deviation

LS Least Squares

LMS Least Mean Squares

LTA Least Trimmed Absolute Deviation

LTS Least Trimmed Squares

M A type of Maximum Likelihood Estimator

MAD Median Absolute Deviation

MCD Minimum Covariance Determinant

MCDreg Minimum Covariance Determinant Regression

MLR Multiple Linear Regression

MVE Minimum Volume Ellipsoid

MLE Maximum Likelihood Estimator

MLTSreg Multivariate Least Trimmed Squares Regression

MM A type of Maximum Likelihood Estimator in two phases

MMD Minimum Mahalanobis Distance Algorithm

MSreg Mean Squared Regression

MSres Mean Squared Residuals

MSreg Multivariate Scale Regression

OGK Orthogonalized Gnanadesikan Kettenring Estimator

OLS Ordinary Least Squares

OLSWO Ordinary Least Squares when observations identified as outliers have been removed from the dataset

PDS Positive Definite and Symmetric Matrix

PRESS Prediction Residuals

RD Robust Distances

REWLS Reweighted Least Squares

SSE Sum of Square Error

SLR Simple Linear Regression

SS Sum of Squares

SSreg Sum of Squares Regression

SSres Sum of Squares Residuals

LIST OF SYMBOLS

α Level of significance

$\boldsymbol{\beta}$ Regression parameters

β_j Element of $\boldsymbol{\beta}$ relating to the j^{th} regressor

$\hat{\boldsymbol{\beta}}$ An estimate of $\boldsymbol{\beta}$

$\hat{\beta}_j$ Element of $\hat{\boldsymbol{\beta}}$ relating to the j^{th} regressor

$\mathbf{C}$ Covariance matrix

$\lceil\ \rceil$ Ceil: Rounds up a float to a specified number of decimal places

$\mathbf{C}_\ell$ *MMD*'s covariance matrix

C_m Main cluster

C_τ Minor cluster

C_{hm} Main cluster containing only h observations

$\mathbf{C_J}$ Covariance matrix of $\mathbf{J}^{th}$ subsample

δ cutoff value for a diagnostic statistic

d_i Mahalanobis distance of i^{th} observation

$d_i(.)$ The i^{th} Mahalanobis distance of an argument

$\boldsymbol{\varepsilon}$ Random error matrix

ε_i The i^{th} random error

$f(.)$ Model function

$\forall$ For all

γ Breakdown point, i.e. fraction of outlier that an estimator can tolerate

$\mathbf{H}$ Hat matrix

h Number of observations defining a half dataset of data

h_{ii} The i^{th} diagonal of $\mathbf{H}$

h_{ij} The i^{th} row and the j^{th} column element of $\mathbf{H}$

$\mathbf{H}_y$ The altered hat matrix

$\mathbf{J}$ Subsample size

J_m Initial sample

$\boldsymbol{l}$ Size of minor clusters to be activated

L_1 Least absolute deviation regression loss function

L_2 Least squares regression loss function

$\mathbf{m}$ Mean vector

$\tilde{\mathbf{m}}$ Median vector

$\mathbf{m}_\ell$ *MMD*'s mean vector

m Size of initial Sample or Initial subset

$\mathbf{m_J}$ Mean vector of $\mathbf{J}^{th}$ subsample

$\boldsymbol{\mu}$ Population Mean

$\hat{\boldsymbol{\mu}}$ Estimate of $\boldsymbol{\mu}$

$MMD(\boldsymbol{\beta})$ *MMD*-variance

N Population size

n Number of observations

p Number of predictor variables

\# Paired sample with parallel and equal size

q Number of response variables

$Rstudent_i$ The Studentized t-like diagnostics

$\mathbf{r}$ Residual vector

r_i Residual for the i^{th} observations

$r_{i,-i}$ Residual values at $\mathbf{x}_i^{'}$ when i^{th} observation is removed

$\boldsymbol{\sigma}^2$ Vector of variances

$\boldsymbol{\sigma}$ Vector of standard deviations

$\boldsymbol{\Sigma}$ Population covariance matrix

$\hat{\boldsymbol{\Sigma}}$ Estimate of $\boldsymbol{\Sigma}$

s Root mean square error for *OLS* regression

s_{-i} Root mean square error for *OLS* regression when i^{th} observation is removed

u_i Robust projection distances

$\mathbf{v}$ Direction vector of $k \times 1$ or $p \times 1$ dimension

x Predictor variable of interest

x_i The i^{th} predictor variable

x_{ji} The i^{th} observation for the j^{th} predictor

$\mathbf{X}$ Predictor matrix including intercept

$\mathbf{x}_i^{'}$ The i^{th} row of $\mathbf{X}$

$\mathbf{X}_y$ Predictor matrix with intercept and augmented with the response vector

$\mathbf{x}_{y,i}^{'}$ The i^{th} row of $\mathbf{X}_y$

y Response variable of interest

y_i The i^{th} response variable

$y_{i,-i}$ Fitted values at $\mathbf{x}_i^{'}$ when i^{th} observation is removed

y_{ki} The i^{th} observation for the k^{th} response

$\hat{\mathbf{y}}$ Predicted response vector

$\mathbf{Z}$ Predictor matrix excluding intercept

$\mathbf{z}_i^{'}$ The i^{th} row of $\mathbf{Z}$

$\mathbf{Z}_y$ Predictor matrix excluding intercept and augmented with the response vector

LIST OF PUBLICATIONS

LIST OF PUBLICATIONS

[1] Alih, E., and Ong, H. C., (2015), Robust Cluster-Based Multivariate Outlier Diagnostics and Parameter Estimation in Regression Analysis, *Communications in Statistics - Simulation and Computation*, Accepted: DOI 10.1080/036 10918.2014.960093. [ISI Impact Factor:0.288]

[2] Alih, E., and Ong, H. C., (2015), An Outlier-resistant Test for Heteroscedasticity in Linear Models, *Journal of Applied Statistics*, **42**:8, 1617-1634. [ISI Impact Factor: 0.656].

[3] Alih, E., and Ong, H. C., (2015), Cluster-based Multivariate outlier identification and Re-weighted Regression in Linear Models, *Journal of Applied Statistics*, **42**:5, 938-955. [ISI Impact Factor: 0.656].

[4] Alih, E., and Ong, H. C., An Outlier resistant algorithm of T^2 Control Chart for Skewed Population, *Statistical Methods & Applications*, Under Review. [ISI Impact Factor: 0.571].

[5] Alih, E., and Ong, H. C., Cluster-based L2 Re-weighted Regression, *Statistical Methodology*, Accepted: DOI 10.1016/j.stamet.2015.05.005. [ISI Impact Factor: 0.708].

[6] Ong, H. C., and Alih, E., Cluster-based Multivariate Regression, *Pakistan Journal of Statistics*, Accepted. [ISI Impact Factor: 0.336].

[7] Ong, H. C., and Alih, E.,(2015), A Control Chart Based on Cluster-Regression Adjustment for Retrospective of Monitoring Individual Characteristics, *PLoS One*, **10**(4):e0125835. [ISI Impact Factor: 3.534].

[8] Alih, E., and Ong, H. C., (2014), The performance of robust multivariate regression in simultaneous dependence of variables in linear models, *AIP Conference Proceedings*, **1606**, pp 1028-1033.

[9] Alih, E., and Ong, H. C., (2015), Cluster-based Control Chart Method, *National Symposium on Mathematical Sciences (SKSM22) Shah Alam*, Kuala Lumpur.

PENGANGGAR BERASASKAN KELOMPOK BAGI MODEL REGRESI BERGANDA DAN LINEAR MULTIVARIAT

ABSTRAK

Dalam bidang pemodelan regresi linear, regresi kuasa dua terkecil (LS) klasik adalah mudah dipengaruhi oleh titik terpencil manakala penganggar regresi rendah-kerosakan seperti regresi M dan regresi pengaruh terbatas mampu menahan pengaruh peratusan kecil titik terpencil. Penganggar tinggi-kerosakan seperti kuasa dua trim terkecil (LTS) dan penganggar regresi (MM) adalah teguh terhadap sebanyak 50% daripada pencemaran data. Masalah prosedur penganggar ini termasuklah permintaan pengkomputeran luas dan kebolehubahan subpensampelan, kerentanan koefisien teruk terhadap kebolehubahan kecil dalam nilai awal, sisihan dalaman daripada trend umum dan kebolehan dalam data bersih dan situasi rendah-kerosakan. Kajian ini mencadangkan suatu penganggar regresi baru yang menyelesaikan masalah dalam model regresi berganda dan regresi multivariat serta menyediakan maklumat berguna tentang kehadiran dan struktur titik terpencil multivariat. Dalam model regresi berganda, prosedur yang dicadangkan menyeragamkan fasa langkah tumpuan (langkah-C) dengan fasa regresi berjujukan. Jarak varians Mahalanobis yang minimum dirujuk sebagai MMD-algoritma tumpuan varians yang menghasilkan penganggar awal. Selepas itu, satu analisis kelompok berhierarki dilakukan dan data seterusnya disusun ke dalam kelompok utama 'set setengah' dan satu kumpulan atau lebih kelompok minor. Suatu anggaran regresi

awal kuasa dua terkecil dihasilkan dari kelompok utama dengan perbezaan dalam statistik sesuai yang mengaktifkan secara jujukan kelompok minor dalam senario regresi pengaruh terbatas. Dalam penetapan regresi multivariat, suatu penentu kovarians jarak minimum Mahalanobis dirujuk sebagai MMCD-algoritma tumpuan kovarians menghasilkan penganggar awal. Jarak reja dikira daripada penganggar awal yang menjadi metrik jarak bagi analisis kelompok berhierarki aglomeratif (AHC). AHC menyusun data seterusnya ke dalam kelompok utama 'set setengah' dan kelompok minor dalam satu kumpulan atau lebih. Anggaran kuasa dua terkecil awal diperolehi daripada kelompok utama. Anggaran awal kemudiannya dioptimumkan dengan menggunakan langkah tumpuan yang menurunkan fungsi objektif reja disetiap langkah tumpuan. Bagi menambahbaik kecekapan anggaran awal, statistik DFFITS digunakan bagi mengaktifkan kelompok minor. Oleh kerana langkah yang dicadangkan bercampur fasa kelompok dengan ulangan fasa regresi kuasa dua terkecil ia dikenali sebagai penganggar berasaskan kelompok bagi model Regresi Berganda dan Linear Multivariat (singkatannya CLreg). CLreg mencapai titik kerosakan-tinggi yang dapat ditentukan oleh pengguna. Ia mewarisi sifat normal asimptot regresi kuasa dua terkecil dan juga varians samaan. Perbandingan kajian kes dan eksperimen simulasi Monte Carlo menunjukkan kelebihan prestasi berbanding cara kerosakan-tinggi lain daripada segi kestabilan koefisien. Satu plot dendogram yang diperolehi daripada analisis kelompok digunakan bagi pengenalpastian titik terpencil multivariat. Secara keseluruhannya, prosedur yang dicadangkan merupakan suatu sumbangan dalam bidang regresi teguh yang menawarkan sudut pandangan berbeza terhadap analisis data dan gabungan antara anggaran dan ringkasan diagnostik.

CLUSTER-BASED ESTIMATORS FOR MULTIPLE AND MULTIVARIATE LINEAR REGRESSION MODELS

ABSTRACT

In the field of linear regression modelling, the classical least squares (LS) regression is susceptible to a single outlier whereas low-breakdown regression estimators like M regression and bounded influence regression are able to resist the influence of a small percentage of outliers. High-breakdown estimators like the least trimmed squares (LTS) and MM regression estimators are resistant to as much as 50% of data contamination. The problems with these estimation procedures include enormous computational demands and subsampling variability, severe coefficient susceptibility to very small variability in initial values, internal deviation from the general trend and capabilities in clean data and in low breakdown situations. This study proposes a new high breakdown regression estimator that addresses these problems in multiple regression and multivariate regression models as well as providing insightful information about the presence and structure of multivariate outliers. In the multiple regression model, the proposed procedures unify a concentration step (C-step) phase with a sequential regression phase. A minimum Mahalanobis distance variance referred to as (MMD)-variance concentration algorithm produces a preliminary estimator. Thereafter, a hierarchical cluster analysis is performed and then the data is partitioned into a main cluster of "half set" and a minor cluster of one or more groups. An initial least squares regression es-

timate arises from the main cluster with a difference in fit statistic ($DFFITS$-statistic) that sequentially activates the minor clusters in a bounded influence regression scenario. In the multivariate regression setting, a minimum Mahalanobis distance covariance determinant referred to as ($MMCD$)-covariance concentration algorithm produces a preliminary estimator. Residual distances computed from this preliminary estimator serves as a distance metric for agglomerative hierarchical cluster (AHC) analysis. The AHC then partition the data into a main cluster of "half set" and a minor cluster of one or more groups. An initial least squares estimate is obtained from the main cluster. The initial estimate is thereafter, optimized using concentration steps that lower the objective function of the residuals at each concentration step. To improve the efficiency of the initial estimates, a difference in fit statistic ($DFFITS$-statistic) is used to activate the minor cluster. Since the proposed method blends the cluster phase with repeated least squares regression phase, it is called the Cluster-based estimators for Multiple and Multivariate Linear regression Models (CLreg for short). CLreg achieves a high breakdown point which can be determined by the user. It inherits the asymptotic normal properties of the least squares regression and is also equivariant. Case studies comparisons and Monte Carlo simulation experiments depict the performance advantage of CLreg over the other high breakdown methods for coefficient stability. A dendrogram plot obtained from cluster analysis is used for multivariate outlier detection. Overall, the proposed procedure is a contribution in the area of robust regression, offering a distinct philosophical viewpoint towards data analysis and the marriage between estimation and diagnostic summary.

CHAPTER 1

INTRODUCTION

1.1 General Introduction

While least squares methods have dominated the statistical literature on regression for many years, a significant interest in alternative methods has emanated in the last few years. One reason for the interest is an increasing awareness of and sensitivity to the problems that occur with the naive application of least squares. In the professional statistical domain, there is a 'usual tale' that the world is fundamentally linear and normal. Interestingly, the 'usual tale' translated into two assumptions viz a viz linearity and normality assumptions. These two assumptions form the bedrock of many statistical procedures over the wide range of fields where statistics is applied. Under these assumptions, there exist in abundance, of attractive statistical theory that is carefully put together into a broad weapon store that can be exploited to analyse data. It may even propose to the analyst, the experimental design pattern needed to optimize a given criteria. This is an outstanding prospect to analysts and researchers, but may as well, produce misleading inferences, or assurances, in the desired output.

Experimental data are usually contaminated data. According to Hampel et al. (2011), real data are presumed to contain between 1% to 10% gross contamination. In regression modelling that studies the functional relationship among variables, the bulk of the data may assume a linear pattern, but contaminated or spurious data points may be

inevitable. The resulting effect is that such contaminated data points may indicate that the regression model is misspecified, that the presumed linear model is not suitable, or better still, human error occurred during experimental stage or in data recording stage. In the long run, a statistical analyst is left at the mercy of the validity of the assumptions being made when conducting analysis. Conditioning a linear model for an experiment as well as restricting outcomes to an independently and identically distributed random error is an assumption of convenience because then, the accompanying analysis to be performed is fixed. On the other hand, assumptions of convenience may not essentially be the valid assumption. Moreover, adapting a given analytical approach exclusively because it is very 'popular' will arbitrarily restrict the quality of desired results.

The advent of computer science has been the reason for a kind of breakthrough in the field of Applied Statistics. Due to the discovery of the personal computer in the 1970s, applications involving computer technology have been growing upwards at a geometric rate. Entree to computer programs and softwares has become almost trivial. Procedures and approximation algorithms that were once seen as too complex to compute are now viable alternatives. Computational confidence is now developed to emphatically pursue new techniques and new algorithms in applied statistics. Statistical procedures that rely on the assumption of Gaussian paradigm such as the classical methodologies are no longer the only viable option. To an extent, convenience has been overridden by technology (Bhatt, 2006).

1.2 Motivation of the Study

Identification of influential observation often referred to as outliers is a common goal of a data analyst. Classical methods using euclidean and Mahalanobis distances have been developed to detect such spurious observations but they fail to detect them because they are affected by the observations they are supposed to identify. Billor et al. (2000) presented an algorithm for selecting the initial subset in a forward search procedure for multivariate outlier identification. Fan et al. (2013) used the hierarchical clustering procedure to improve the capability of certain multivariate control chart methods in identifying the presence of multivariate outliers. Hubert et al. (2012) proposed the deterministic algorithm for robust location and scatter. It turns out that the robust Mahalanobis distance computed from the robust location and scatter matrix forms the basis for outlier identification. The problem with these estimators is that they are impractical to compute exactly in large samples. As a result, approximation algorithms are used. The algorithms generally produce estimators with lower consistency rates and breakdown values than the exact theoretical estimator. This discrepancy grows sporadically with sample size, with the implication that huge computations are required for good approximations in large high-dimensional samples. In the end, masking and swamping effects lead to false identification of observation as either outlying or otherwise. This phenomenon motivated this study to investigate and propose an alternative outlier identification procedure.

The classical least squares (LS) regression produces the 'best' estimates when assumptions of Gaussian paradigm are all valid. However, the presence of outliers influence distort these estimates so much so that their values are no longer reliable. The gener-

alized regression M-estimator reported in Huber and Ronchetti (2011) is an example of robust regression alternative for small percentage of outliers in the y-dimension. Hamzah and Nasser (2014) discussed various types of high breakdown M-regression estimation in the context of generalized linear models; while the least median of squares (LMS) estimator reported in Rousseeuw and Leroy (2003), least trimmed squares (LTS) estimator reported in Rousseeuw and Leroy (2003) and the MM estimator reported in Maronna et al. (2006) are examples of high breakdown regression estimators that can resist up to 50% contaminations in datasets. The problem with these robust alternatives is that the generalized regression M-estimator can only handle small percentage of outliers in the y-dimension and breaks down when x-outliers are present. According to Hawkins and Olive (2002), the high break down regression estimators such as LMS, LTS and MM on the other hand implement 'elemental set ' or 'resampling algorithms' . These algorithms turn out to be completely ineffective in high dimensions with high level of outlier observations. This is because the algorithm involves combinatoric-based analysis which becomes completely overwhelmed even with modest sample size. The phenomenon injects some sort of random sample variability into the analysis. Hence, reproducibility or lack thereof, becomes a real issue. Accordingly, the need for an alternative robust regression in multiple regression model is desirable.

According to Rousseeuw et al. (2004), the classical multiple regression is highly susceptible to little contaminations in datasets. This setback also occurs in multivariate regression analysis. Methods such as MCD regression, multivariate S-estimator for robust estimation and inferences, and multivariate least trimmed squares regression estimators are robust alternatives to the classical multivariate least squares regression.

These robust alternatives like their multiple regression counterparts implement 'resampling algorithms ' and 'elemental subset'. The drawback is that methods that implement 'resampling algorithms' and 'elemental subset' are ineffective and fail to produce estimators that are reliable in high dimension. They are characterized with poor efficiency and lack of reproducibility. This research is motivated by these ill-effects to investigate and propose an alternative multivariate robust regression estimator to address the inherent issue with existing alternatives.

1.3 Research Objectives

This research is centered on the study of robust, high breakdown estimators for location and dispersion in multivariate outlier identification and its extension to regression modeling in the multiple and multivariate scenarios with the following objectives.

1. To develop a new estimator for outlier identification that is:

   (a) Robust and utilizes substantial information about the sampling distribution,

   (b) Able to handle masking and swamping effect in the presence of extreme observations,

   (c) Not computationally intensive.

2. To develop a new regression procedure in the multiple and multivariate models that:

   (a) Achieve high breakdown regression parameter estimates,

(b) Offers informative summary about likely multivariate outlier structure,

1.4 Thesis Organization

The remaining parts of chapter 1 discuss the background of classical regression approach alongside some classical outlier diagnostic in regression analysis. A case study is presented for the purpose of illustration and comparisons. Chapter 2 deals with literature review of robust methods such as the various robust procedures for computing location and dispersion estimators and outlier identification procedures, followed by a review of robust regression methods with emphasis on low breakdown and high breakdown scenarios for multiple regression. A review of existing robust alternatives to *LS* multivariate regression concludes chapter 2. The proposed method for outlier identification is presented in chapter 3 alongside its equivariance property, Monte Carlo simulation experiment and numerical examples. The proposed algorithm for multiple regression method is presented in chapter 4 alongside its numerical illustrations, Monte Carlo simulation and theoretical properties. Chapter 5 introduces the proposed multivariate regression method, its theoretical robustness, Monte Carlo simulation experiment and real data application. Chapter 6 concludes the research work with summary of findings, and areas for future research.

1.5 Regression background and Notation

Regression study can be seen as the study of how two sets of variables are related. The first set of variables defines the p-predictors denoted as $x_{ji} = x_{1i}, x_{2i}, x_{3i}, \cdots, x_{pi}; i = 1, \cdots, n; j = 1, \cdots, p$. The Gaussian paradigm assumes that these predictor variables

are fixed. They are sometimes called regressor or independent variables. Their numerical values are either experimental outcomes or sometimes designed in advance of the experiment. The second set of variables contains the q-responses denoted as $y_{ki} = y_{1i}, y_{2i}, y_{3i}, \cdots, y_{qi}; i = 1, \cdots, n;\ k = 1, \cdots, q$. They are often called dependent variables since their values depend on the regressor variables. Three scenarios can be described by the aggregation of regressor and response variables. Scenario 1 describes a situation where $p = q = 1$ and is referred to as the simple linear regression model. The second scenario is when $p > 1$, $q = 1$ and it is referred to as the multiple linear regression model. Scenario 3 is when $p > 1, q > 1$ and it is referred to as the multivariate linear regression model.

The generalized statistical model that expresses the response as a function of the regressor variables with the addition of an error term arising from the random experiment is given as

$$y_i = f(x_{1i}, x_{2i}, x_{3i}, \cdots, x_{pi}) + \varepsilon_i, i = 1, \cdots, n \tag{1.1}$$

where n is the sample size. The specific form of the model in Equation (1.1) is restricted to a family of linear function given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \cdots, \beta_p x_{pi} + \varepsilon_i. \tag{1.2}$$

Note that Equation (1.2) is linear in terms of the unknown parameters $\boldsymbol{\beta}_j \in \mathbb{R}^{p\times 1}$.

In line with the typical linear regression notation, the data are displayed in an $n \times 1$ response vector, $\mathbf{y}$, and an $n \times p$ regressor matrix, $\mathbf{X}$. Define $\boldsymbol{\beta}_j \in \mathbb{R}^{p\times 1}$ as the parameter

vector, and ε as the $n \times 1$ error vector, the linear regression model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.3}$$

The model in Equation (1.3) can be written elementwise as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & x_{32} & \cdots & x_{p2} \\ 1 & x_{13} & x_{23} & x_{33} & \cdots & x_{p3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

For easy understanding, it is important to define some matrices that are pivotal in the computational process of various regression methods. Create the $n \times (p+2)$ matrix $\mathbf{X}_y$ by augmenting the vector of $\mathbf{y}$ to the matrix $\mathbf{X}$ to obtain

$$\mathbf{X}_y = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & \cdots & x_{p1} & y_1 \\ 1 & x_{12} & x_{22} & x_{32} & \cdots & x_{p2} & y_2 \\ 1 & x_{13} & x_{23} & x_{33} & \cdots & x_{p3} & y_3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} & \cdots & x_{pn} & y_n \end{bmatrix}.$$

In Chapter 3 where the proposed method for outlier identification will be discussed, as well as location and dispersion estimators, it becomes crucial to remove the column of ones from the design matrix $\mathbf{X}$ and the augmented matrix $\mathbf{X}_y$. Let $\mathbf{Z}$ be the $n \times p$ matrix containing only the p-regressors,

$$\mathbf{Z}=\begin{bmatrix} x_{11} & x_{21} & x_{31} & \cdots & x_{p1} \\ x_{12} & x_{22} & x_{32} & \cdots & x_{p2} \\ x_{13} & x_{23} & x_{33} & \cdots & x_{p3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & x_{3n} & \cdots & x_{pn} \end{bmatrix},$$

such that $\mathbf{Z}_y$ represents the $n \times (p+1)$ matrix formed by augmenting the vector $\mathbf{y}$ to $\mathbf{Z}$, given by

$$\mathbf{Z}_y=\begin{bmatrix} x_{11} & x_{21} & x_{31} & \cdots & x_{p1} & y_1 \\ x_{12} & x_{22} & x_{32} & \cdots & x_{p2} & y_2 \\ x_{13} & x_{23} & x_{33} & \cdots & x_{p3} & y_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{1n} & x_{2n} & x_{3n} & \cdots & x_{pn} & y_n \end{bmatrix}.$$

In order to be able to reference each of the individual observations, the i^{th} row of $\mathbf{X}$ will be referred to as the $1 \times p$ row vector $\mathbf{x}_i'$, and the $1 \times p$ row of $\mathbf{z}_i'$ will represent the i^{th} row of $\mathbf{Z}$. When the dependent variable is augmented, the notation becomes $\mathbf{x}_{y,i}'$ and $\mathbf{z}_{y,i}'$ for the i^{th} row of $\mathbf{X}_y$ and $\mathbf{Z}_y$ respectively.

The 'hat ' symbol is used to denote estimates of parameters. For instance, $\hat{\boldsymbol{\beta}}$ denotes the estimates of the parameter vector $\boldsymbol{\beta}$ and the $n \times 1$ vector of predicted responses is obtained as $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ such that $\hat{y}_i$ represents the predicted response at the i^{th} regressor space. Also, the $n \times 1$ vector of residuals is computed as $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ such that r_i is the residual for the i^{th} observation.

1.6 The Least squares Regression-*LS*

The classical least squares (LS)-regression procedure is a method of estimating the parameter vector $\boldsymbol{\beta}$ of the regression model that is based on the assumption of Gaussian paradigm. Apart from the model specification assumption, the *LS* has other assumptions concerning the error term. It assumes that the errors are independent and identically distributed (iid) random variables whose distribution is that of a normal random variable with mean 0 and a constant variance σ^2. The vector of parameter estimates, $\hat{\boldsymbol{\beta}}$, is computed from the loss function popularly referred to as Ordinary Least Squares (OLS):

$$\min_{\forall\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_{i3} x_{i3} - \cdots - \beta_p x_{ip})^2$$

The expression describes the deviation between the observed response values, y_i and their corresponding predicted response values, $\hat{y}_i$ which is a function of $\mathbf{x}_i'$. Hence, the ordinary least squares (OLS) minimizes the sum of squared residuals, as a result, it is named least squares (LS).

When the assumptions of Gaussian paradigm are all valid, the *LS*-estimator is said to be the 'best' linear unbiased estimator, or *BLUE*. This implies that the *LS*-estimator has the least deviation among all other linear unbiased estimators. Moreover, the maximum likelihood estimator (MLE) turns out to be the *LS*-estimator under the validity of these assumptions.

1.6.1 Terminology

Robust regression deals mainly with the tendency of an estimator to resist the influence of extreme observations in datasets. The terms outliers and leverages are often used to describe these extreme observations, thus, they are defined below.

An outlier can be described as an observation that is extreme in the response space in relation to the general pattern exhibited by the bulk of the data. Leverage on the other hand is the term used to define the location of an observation in the predictor space. A low leverage observation is the data point located near the central tendency of the predictors while a high leverage observation stays in some extreme position far away from the central tendency of the predictors. In a concise term a point $(\mathbf{y}_i', \mathbf{x}_i')$ which does not follow the linear pattern of the bulk of the data but whose $\mathbf{x}_i'$ is not outlying is called vertical outlier. A point $(\mathbf{y}_i', \mathbf{x}_i')$ whose corresponding $\mathbf{x}_i'$ is outlying is called a leverage point. A point $(\mathbf{y}_i', \mathbf{x}_i')$ is a bad leverage point when it does not follow the pattern of the bulk of the data; otherwise, it is a good leverage point (this is because it does not harm the regression fit and icrease efficiency).

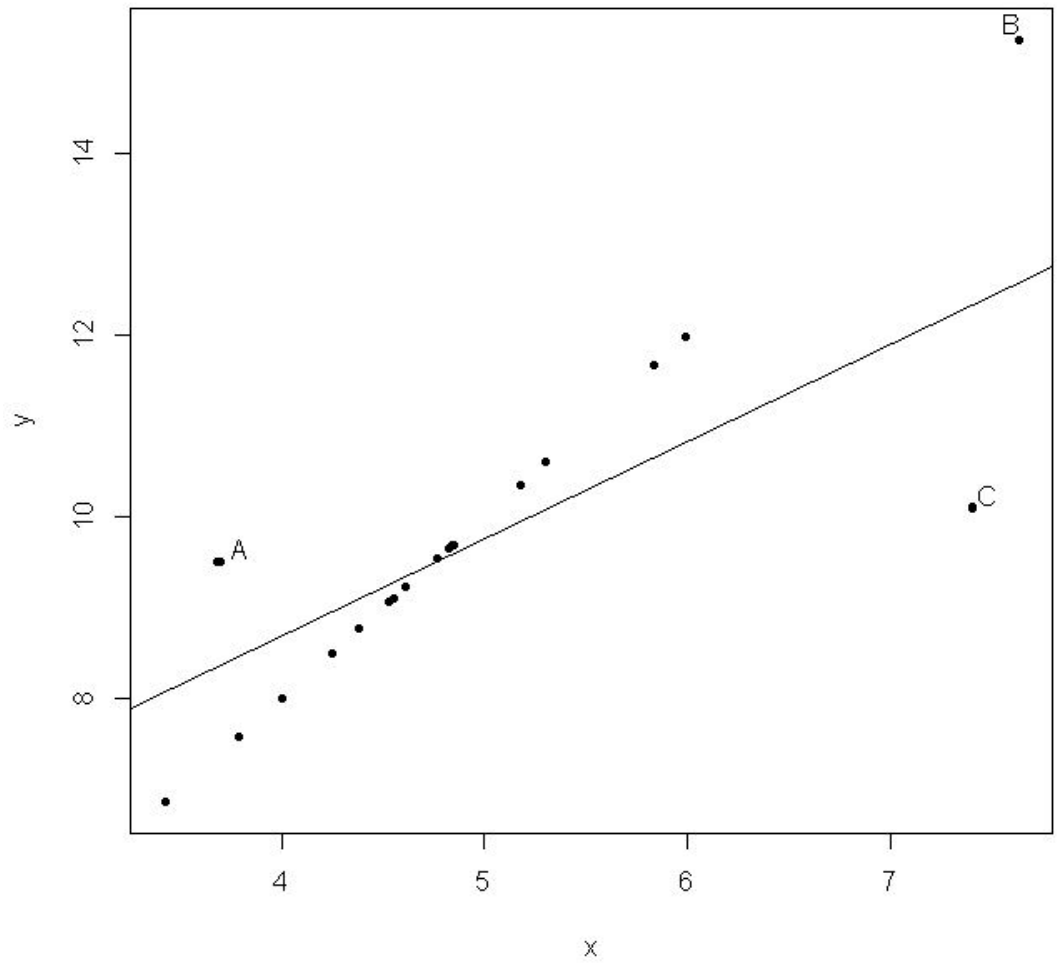

Figure 1.1: Scatter Plot Illustrating Outliers and Leverages

Figure 1.1 depicts the differences between outliers, good leverage points and bad leverage points. Point *A* is an outlier, point *B* is a good leverage observation, and point *C* is a bad leverage observation.

The *LS* regression is extremely susceptible to influential observations. This is because an outlier may have a large residual which, when squared, it wrecks the *LS* objective function. This is illustrated in the *LS* regression fit of Figure 1.1 where points *A* and *C* pulled the *LS* regression line from passing through point *B* where the bulk of the data trend follows.

To summarize, the *LS* regression is susceptible to as little as one influential observation. As a consequence, parameter estimates and inferential statistics about the estimation process turns out to be misleading. If *LS* regression is blindly implemented without any exploratory data analysis, it may be unknown when the analysis has severe flaws. The phenomenon may lead to a conclusion that does not support the data in the actual sense. It is therefore, necessary to make a right choice of regression methodology in data analysis.

With the intent to curtail the drawbacks and risks of failure of Gaussian assumptions in *LS* procedure, diagnostic approaches have been proposed for the identification of contaminated observations. These diagnostic tools are applicable in the framework of robust regression as well.

1.7 Leverage and the Hat Matrix

By definition, the tendency for an observation to overwhelm an estimator relative to its position in the predictor hyperplane, irrespective of whether the corresponding response variable is an outlier or not is referred to as an observation's leverage. The major difference between conventional regression and robust regression lies in the quantification of this concept.

To begin with, for a p-dimensional predictor variables in a regression model without intercept parameter, the *Mahalanobis distance metric*, d_i, for the i^{th} observation is given by

$$d_i = \sqrt{(\mathbf{z}_i - \mathbf{m})' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{m})}, \tag{1.4}$$

where $\mathbf{m}$ is the $p \times 1$ mean vector and $\mathbf{C}$ is the $p \times p$ covariance matrix.

Furthermore, consider the scenario in which the intercept parameter is included in the model, the '*hat matrix*' replaces the Mahalanobis distances. This matrix (*hat matrix*) described as the $p \times p$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, acquired the name *hat matrix* because it maps the observed response variable into the predicted values of the response variables as $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{y}$. This matrix plays an important role in classical outlier diagnostics. Myers (2000) enumerated the properties of the *Hat matrix* as follows:

1. The hat matrix is symmetric and idempotent.

2. The trace of $\mathbf{H}$ equals p, the dimensional size: i.e $\sum_{i=1}^{n} h_{ii} = p$.

3. The i^{th} hat diagonal h_{ii} is bounded such that:

   (a) For models with intercept, $\frac{1}{n} \leq h_{ii} \leq 1, \forall i$,

   (b) For models without intercept, $0 \leq h_{ii} \leq 1, \forall i$.

4. If $h_{ii} = 1$, then $h_{ij} = h_{ji} = 0, \forall i \neq j$.

5. Every row of $\mathbf{H}$ (and by symmetry, every column) sums to one i.e. $\sum_{j=1}^{n} h_{ij} = 1, \forall i$.

6. The i^{th} hat diagonal, h_{ii} is monotonically related to the i^{th} Mahalanobis distance d_i such

$$h_{ii} = \frac{d_i^2}{n-1} + \frac{1}{n}. \tag{1.5}$$

The hat diagonal primarily measures how far an observation is from the central ten-

dency in the predictor space (Myers, 2000). As a result, the *LS*-regression procedure uses the hat diagonal to measure its leverage. Since the sum of all n hat diagonals equals p, and thus $\frac{2p}{n}$ is twice the mean hat diagonal, Myers (2000) proposed a rule of thumb for identifying leverage points as any observation with $h_{ii} \geq \frac{2p}{n}$.

The main drawback of the hat diagonals and the Mahalanobis distance is that they are both functions of the mean. As a result they are affected by the observation they are supposed to identify. As can be seen from Equation (1.5), describing the relationship of hat diagonal to the Mahalanobis distances, these hat diagonals are functions of the classical mean denoted as **m**, and the corresponding sample covariance denoted as **C**. According to Billor et al. (2000), both **m** and **C** are susceptible to influential observations, and hence, they are not reliable estimators. Consequently, the discussion of more resistant estimator for **m** and **C** is hitherto a desirable alternative. Precisely, a robust Mahalanobis distance metric will be proposed by substituting **m** and **C** by more resistant estimators of multivariate location and dispersion.

1.7.1 Altered Hat Matrix

The hat matrix is based on the predictor variables only. As a result, it does not account for outlying observations in the **y**-subspace. In order to account for the likely presence of outliers in the response variable, the $n \times n$ altered hat matrix, $\mathbf{H}_y$ is described as $\mathbf{H}_y = \mathbf{X}_y(\mathbf{X}_y'\mathbf{X}_y)^{-1}\mathbf{X}_y'$. Myers (2000) reported that the altered hat matrix also have the properties of the hat matrix enumerated in Section 1.7. Its trace is now $p+1$, as against

p for the usual hat matrix. Myers (2000) has shown that

$$\mathbf{H}_y = \mathbf{H} + \frac{\mathbf{rr}'}{\mathbf{SSE}}, \tag{1.6}$$

where $\mathbf{r}$ is the residual vector and **SSE** the sum of squares error, both computed from the *LS*-regression. Since $\mathbf{H}_y$ is based on $\mathbf{H}$, *LS* residuals and **SSE**, its diagonal elements are susceptible to outliers or high leverage points.

1.8 Outlier Diagnostics

In searching for outliers in the response variable, it is ideal to begin with residual analysis. However, Myers (2000) stated that high leverage points tend to have small residuals because they overwhelm the fitted values and are masked by it. Consequently, a cautionary approach when making inferences from the set of residuals is ideal.

Given that residuals constitute the units of the response variable, the extent of what constitutes a large residual is a function of the scale parameter. Hence, an estimate for σ, denoted by S is used to re-scale the residuals. Myers (2000) suggested the root mean square error $(RMSE)$ obtained from the classical regression analysis of variance $(ANOVA)$ table. The *RMSE* is then, defined in terms of an *internally studentized residual* given by

$$r_i^{'} = \frac{r_i}{s\sqrt{1-h_{ii}}}. \tag{1.7}$$

where r_i is the i^{th} element of the residual vector $\mathbf{r}$. Lest for lack of independence, Equation (1.7) follows a Student t-distribution. Given that the numerator and denominator are not independent and $r_i^{'}$ does not follow a true t-distribution. However, Myers

(2000) deemed it to be nearly t-like, and suggested that an i^{th} observation is deemed outlier if the corresponding $|r_i^{'}| \geq 2$.

The performance of a regression procedure can be assessed by viewing the regression analysis when a particular observation is included and when it is removed and then measure the change in both scenarios. This type of assessment is called 'single point deletion analysis' or the 'prediction residual $(PRESS)$ approach'. For purpose of notation, '$-i$' subscript could mean that the analysis is carried out without the i^{th} data point. Therefore, $\hat{y}_{i,-i}$ denotes the fitted value at $\mathbf{x}_i'$ when the regression is carried out without $\mathbf{x}_i'$ and y_i. Ordinarily, the *PRESS* analysis appears to involve the computation of n separate regressions, each with a reduced sample size $n-1$. As a result of manipulations, only the full data regression is required. That is, the residual $r_{i,-i} = y_i - \hat{y}_{i,-i}$ usually referred to as the *PRESS* residual can be computed as $r_{i,-i} = \frac{r_i}{(1-h_{ii})}$.

The presence of outlier in the response space tends to inflate the estimate of scale, s^2 . To deflate such inflation, an alternative to the internally studentized residual is desirable. Myers (2000) suggested to compute, s^2_{-i}, in the *PRESS* procedure, where

$$s_{-i} = \sqrt{\frac{(n-p)s^2 - r_i^2/(1-h_{ii})}{n-p-1}} \tag{1.8}$$

is used to compute the *externally studentized residuals* defined as

$$Rstudent_i = \frac{y_i - \hat{y}_i}{s_{-i}\sqrt{1-h_{ii}}}. \tag{1.9}$$

In the opinion of Myers (2000) the Rstudent-statistic follows a t-distribution with $n-p-1$ degrees of freedom, denoted by t_{n-p-1}. Under the Gaussian assumptions, the

numerator and denominator are now independent. An outlier is identified by a Rstudent value if $|Rstudent_i| > t_{(1-\alpha),(n-p-1)}$. When the effect of the i^{th} data point is high, the data point will have a larger Rstudent value than the internally studentized residual. This is because the estimate of scale s^2_{-i} used in the denominator without this data point tends to be smaller than the estimate of scale obtained when using all of the data.

1.9 Influence Diagnostics

Huber (1981) defines the term *influence*, as the relative effect of a particular observation's presence on the resulting estimator. Since high leverage outlier has a distorting effect on the *LS* estimator, their deletion has a significant effect on the estimator. Therefore, they are said to be substantially influential. Armed with this definition, one can infer that single point deletion analysis plays a significant role in influence diagnostics. The influence of the i^{th} observation can be measured using the statistic referred to as 'difference in fits' (DFFITS)-statistic, given as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}}. \tag{1.10}$$

Myers (2000) have shown that the $DFFITS$-statistic can be written as

$$DFFITS_i = Rstudent_i[\frac{h_{ii}}{1-h_{ii}}]^{\frac{1}{2}} \tag{1.11}$$

The advantage of the $DFFITS$-statistic as an influential diagnostic tool is that it constitutes an outlier diagnostic component, $Rstudent_i$, and a leverage diagnostic component, $[\frac{h_{ii}}{(1-h_{ii})}]^{\frac{1}{2}}$. Hence, the $DFFITS$-statistic seems to be a better alternative for both

the outliers and leverages in its evaluation of an observation's influence on an estimator (Atkinson et al., 2004). For this reason, the $DFFITS$-statistic is used as a measure to determine the influence of an outlier cluster on the proposed multiple regression model.

More robust outlier identification, location and dispersion estimator as well as outlier-resistant regression methodologies are now discussed in Chapter 2. These estimators are resistant to contaminations of various sorts and can work well in data analysis when outliers are indeed present, but go unidentified by the conventional procedure.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter discusses the various robust procedures for computing location and dispersion estimators and outlier identification procedures. In what follows is the review of robust regression methods with emphasis on low breakdown and high breakdown scenarios for multiple regression. A review of existing robust alternatives to *LS* multivariate regression concludes this chapter.

Traditionally, nearly all robust outlier identification procedures are based on the location and dispersion estimators. This accounts for why the two (location and dispersion estimator and outlier identification methods) are jointly discussed. Outlier identification methods can be broadly classified into two groups, namely, the forward search algorithms and the elemental or resampling set algorithms. The forward search procedure heuristically searches for a 'half subset' from a data set through a forward stepping algorithm. The method involves selecting an initial subset that is outlier-free and then load it (initial subset) according to a decision criterion until the initial subset contains 'half subset'. Observations with indexes corresponding to the half subset is used to compute the robust location and dispersion. The robust location and dispersion estimators are in turn, used to compute the Mahalanobis distance. Observations whose robust Mahalanobis distance is greater than a specified threshold (say δ) is declared

as an outlier. The coordinatewise median, the Stahel-Donoho estimator, the Stalactite plot analysis, the Orthogonalized Gnanadesikan-Kettenring estimator (OGK) and the block adaptive computationally efficient outlier nominator $(BACON)$ are examples of the forward search algorithm procedure. The elemental or resampling set procedures on the other hand computes the location and dispersion estimators by taking a subset from a data set based on a given objective function that minimizes the error in the subset selection criterion. A resampling step then optimizes this criterion by lowering the objective function at each resampling step until an optimum subset which is outlier-free is obtained. A location and dispersion estimator arising from this outlier-free subset is then used to compute a Mahalanobis distance. A candidate outlier is that observation whose index corresponds to the Mahalanobis distance that is greater than a specified threshold (say δ). The minimum volume ellipsoid (MVE), the minimum covariance determinant (MCD) and the Deterministic minimum covariance determinant $(DetMCD)$ are members of robust outlier identification that implement the elemental or resampling procedure.

The goal of robust statistics in outlier identification is the desire to obtain a robust multivariate location estimator, denoted by $\mathbf{m}$, and a multivariate dispersion estimator, also denoted as $\mathbf{C}$ in such a way that the influence of spurious observations can be subdued. The significance of these techniques to robust regression arena hinges on the ability to extract information about outlier presence as well as its structures especially in multivariate settings. This implies a transformation from the Gaussian approach of 'equal weight assignment' into another optimistic more robust approach of weight assignment to observations based on their influence in the data space. It is important to state at this point that any mention of 'half set' or 'half dataset' corresponds to $h = [(n+$

$p+1)/2]$-observations in outlier identification as well as multiple regression settings. It is modified as $h = [(n+k+1)/2]; k = p+q$ in multivariate regression settings. The outlier identification procedures based on the location and dispersion estimators discussed below is relative to the data matrix $\mathbf{Z}_y$.

2.2 Location and Dispersion Estimators

Suppose that $\mathbf{Z}_y$ is taken from a population with a multivariate normal distribution, then the classical mean and covariance estimators are computed, respectively, as the $k \times 1$ sample mean vector,

$$\mathbf{m} = \frac{\sum_{i=1}^{n} \mathbf{z}_{y,i}}{n} \tag{2.1}$$

and the $k \times k$ sample covariance (dispersion) matrix,

$$\mathbf{C} = \frac{\sum_{i=1}^{n} (\mathbf{z}_{y,i} - \mathbf{m})(\mathbf{z}_{y,i} - \mathbf{m})'}{n-1} \tag{2.2}$$

Equations (2.1) and (2.2) are apparently, functions of the mean and hence, any outlier observation can erratically deflate or inflate them. The repercussions are that inferences drawn from these estimators, such as the hat diagonals, for instance, are severely unreliable in the presence of unusual or extreme observations. Therefore, it is imperative to think of alternative estimators that can resist the influence of these extreme observations.

One simple way to estimate $\mathbf{m}$ and $\mathbf{C}$ is to adopt a robust multivariate location estimator in a coordinatewise manner. In the same way as the univariate location estima-

tion, replace the sample mean by the sample median for each of the k-variables. The covariance matrix turns out to be the covariance estimation that is centered by this coordinatewise median. According to Billor et al. (2000), the coordinatewise median is thought to be more robust as much as 50% than the mean. Although the coordinatewise median and the dispersion estimator computed from it is robust, computational convenience appears to have determined the choice of the median in the past. This is because in the multivariate setting, the coordinatewise median location estimator is not affine equivariant. According to Rousseeuw and Leroy (2003), this estimator is not affine equivariant, implying that linear transformations of the data are not equivalently transformed by the estimator. The procedures discussed below combine outlier identification with estimation of location and dispersion estimators. They are classified into: the forward search procedures and elemental or resampling procedures.

2.2.1 The Forward Search Procedures

Several estimators exist that can replace the non robust sample mean vector and dispersion matrix estimators. Some of these estimators differ mainly in terms of their objective functions and theoretical properties. The following are major outlier resistant estimators for multivariate location and dispersion that adopt the forward search procedure.

2.2.1(a) The Stahel-Donoho Estimator

Proposed independently by Stahel (1981), Donoho (1982) and Dodge (2002) described the Stahel-Donoho estimator for outlier identification and location and dispersion estimation. In simple words, their notion is that a spurious data point will cluster dif-

ferently from the bulk of the data when they are outlying in a correct metrics. The procedure is implemented in two stages. A projection computation is used to determine a robust distance in the first stage. At the second stage, the robust distances serve as a weight function with a weighted mean vector and weighted dispersion matrix. According to Rousseeuw and Leroy (1987), this approach is affine equivariant, and attains high breakdown point when the data are in general position. The algorithm itself is presented below.

Algorithm 2.1: The Stahel-Donoho Projection-based Location and Dispersion Algorithm; *Source:* Rousseeuw and Leroy (1987)

1. Obtain a robust distance, denoted by u_i, for the i^{th} observation $\mathbf{z}_{y,i}$, as:

$$u_i = \sup_{||\mathbf{v}||=1} \frac{|\mathbf{z}_{y,i}^{'}\mathbf{v} - \underset{\forall j}{med}(\mathbf{z}_{y,j}^{'}\mathbf{v})|}{\underset{\forall k}{med}|\mathbf{z}_{y,k}^{'}\mathbf{v} - \underset{\forall j}{med}(\mathbf{z}_{y,j}^{'}\mathbf{v})|}, \tag{2.3}$$

   where $\mathbf{v}$ is the $k \times 1$ directional vector, through which the projections of all n observations are made. The robust distance that is a function of weights obtained from u_i is used to classify outliers into a different cluster while the clean data points remain in the main cluster.

2. Using the n robust distances, u_i and the data set $\mathbf{Z}_y$ the robust location estimator is computed as

$$\mathbf{m} = \frac{\sum_{i=1}^{n} w(u_i)\mathbf{z}_{y,i}}{\sum_{i=1}^{n} w(u_i)}, \tag{2.4}$$

   and the weighted covariance matrix is computed based on the location estimator