UNIVERSITI PUTRA MALAYSIA

ROBUST TECHNIQUES FOR LINEAR REGRESSION WITH MULTICOLLINEARITY AND OUTLIERS

MOHAMMED ABDULHUSSEIN MOHAMMED

IPM 2016 1

ROBUST TECHNIQUES FOR LINEAR REGRESSION WITH MULTICOLLINEARITY AND OUTLIERS

By

MOHAMMED ABDULHUSSEIN MOHAMMED

Thesis Submitted to the School of Graduated Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Degree of Doctor of Philosophy

January 2016

DEDICATIONS

To the spirit of my father and my beloved mother

To my wife for all his contribution, patience and understanding throughout my doctoral studies. He incredibly supported me and made it all possible for me.

To my son, Taha, and my daughters Ethar and Kawther who was accompanying me in all different parts of my study and his love has always been my greatest inspiration

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirements for the degree of Doctor of Philosophy

ROBUST TECHNIQUES FOR LINEAR REGRESSION WITH MULTICOLLINEARITY AND OUTLIERS

By

MOHAMMED ABDULHUSSEIN MOHAMMED

January 2016

Chair : Professor Habshah Midi, PhD
Faculty : Institute for Mathematical Research

The ordinary least squares (OLS) method is the most commonly used method in multiple linear regression model due to its optimal properties and ease of computation. Unfortunately, in the presence of multicollinearity and outlying observations in a data set, the OLS estimate is inefficient with inflated standard errors. Outlying observations can be classified into different types, such as vertical outlier, high leverage points (HLPs) and influential observations (IO).

It is very crucial to identify HLPs and IO because of their responsibility for having large effect on various estimators, causing masking and swamping of outliers in multiple linear regression. All the commonly used diagnostic measures fail to correctly identify those observations. Hence, a new improvised diagnostic robust generalized potential (IDRGP) is proposed. The proposed IDRGP is very successful in detecting multiple HLPs with smaller masking and swamping rates.

This thesis also concerned on the diagnostic measures for the identification of bad influential observations (BIO). The detection of BIO is very important because it is accountable for inaccurate prediction and invalid inferential statements as it has large impact on the computed values of various estimates. The Generalized version of DFFITS (GDFF) was developed only to identify IO without taking into consideration whether it is good or bad influential observations. In addition, although GDFF can detect multiple IO, it has a tendency to detect lesser IO as it should be due to swamping and masking effect. A new proposed method which is called the modified generalized DFFITS (MGDFF) is developed in this regard, whereby the suspected HLPs in the initial subset are identified using our proposed IDRGP diagnostic method.

To the best of our knowledge, no research is done on the classification of observations into regular, good and bad IOs. Thus, the IDRGP-MGDFF plot is formulated to close the gap in the literature.

This thesis also addresses the issue of multicollinearity problem in multiple linear regression models with regards to two sources. The first source is due to HLPs and the second source of multicollinearity problem is caused by the data collection method employed, constraints on the model or in the population, model specification and an over defined model. However, no research is focused on the parameter estimation method to remedy the problem of multicollinearity which is due to multiple HLPs. Hence, we propose a new estimation method namely the modified GM-estimator (MGM) based on MGDFF. The results of the study indicate that the MGM estimator is the most efficient method to rectify the problem of multicollinearity which is caused by HLPs.

When multicollinearity is due to other sources (not HLPs), several classical methods are available. Among them, the Ridge Regression (RR), Jackknife Ridge Regression (JRR) and Latent Root Regression (LRR) are put forward to remedy this problem. Nevertheless, it is now evident that these classical estimation methods perform poorly when outliers exist in a data. In this regard, we propose two types of robust estimation methods. The first type is an improved version of the LRR to rectify the simultaneous problems of multicollinearity and outliers. The proposed method is formulated by incorporating robust MM-estimator and the modified generalized M-estimator (MGM) in the LRR algorithm. We call these methods the Latent Root MM-based (LRMMB) and the Latent Root MGM-based (LRMGMB) methods.

Similar to the first type, the second type of robust multicollinearity estimation method also aims to improve the performance of the robust jackknife ridge regression. The MM-estimator and the MGM-estimator are integrated in the JRR algorithm for the establishment of the improved versions of JRR. The suggested method is called jackknife ridge MM-based denoted by JRMMB and the jackknife ridge MGM based denoted by JRMGMB. All the proposed methods outperform the commonly used methods when multicollinearity comes together with the existence of multiple HLPs.

The classical multicollinearity diagnostic measure is not suitable to correctly diagnose the existence of multicollinearity in the presence of multiple HLPs. When the classical VIF is employed, HLPs will be responsible for the increased and decreased of multicollinearity pattern. This will give misleading conclusion and incorrect indicator for solving multicollinearity problem. In this respect, we propose robust VIF denoted as RVIF(JACK-MGM) which serves as good indicator that can help statistics practitioners to choose appropriate estimator to solve multicollinearity problem.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

TEKNIK TEGUH BAGI REGRESI LINEAR DENGAN MULTIKOLINEARAN DAN TITIK TERPENCIL

Oleh

MOHAMMED ABDULHUSSEIN MOHAMMED

Januari 2016

Pengerusi : Profesor Habshah Midi, PhD
Fakult : Institut Penyelidikan Matematik

Kaedah biasa kuasa dua terkecil (OLS) merupakan kaedah yang sering digunapakai dalam model regresi linear berganda kerana ia mempunyai ciri-ciri optimum dan pengiraan yang mudah. Malangnya, dengan kehadiran multikolinearan dan titik terpencil dalam data, penganggar OLS menjadi tidak cekap dengan ralat piawai yang tinggi. Cerapan titik terpencil boleh dikelaskan kepada pelbagai jenis, iaitu titik terpencil menegak, titik tuasan tinggi (HLP) dan cerapan berpengaruh (IO).

Adalah sangat penting untuk mengenalpasti titik tuasan tinggi dan cerapan berpengaruh kerana keduanya bertangungjawab memberi pengaruh besar keatas pelbagaian penganggar menyebabkan limpahan dan litupan titik terpencil dalam regresi linear berganda. Kesemua pengukuran diagnostik yang biasa digunakan gagal mengenalpasti cerapan tersebut dengan tepat. Oleh itu, kaedah baru Penambahbaikan Berdiagnostik Teguh Potensi Teritlak (IDRGP) dicadangkan. Kaedah IDRGP yang dicadangkan sangat berjaya mengenalpasti kesemua titik tuasan tinggi berganda dengan kadar litupan dan limpahan lebih kecil.

Tesis ini juga mempertimbangkan pengukuran diagnostik bagi mengenalpasti cerapan berpengaruh buruk (BIO). Pengenalpastian BIO amat penting kerana ia penyebab kepada ketidaktepatan ramalan dan inferensi tidak sah disebabkan ia memberi kesan besar keatas nilai pelbagai penganggaran yang dikira. Versi teritlak DFFITS (GDFF) telah dibangunkan hanya untuk mengenalpasti cerapan berpengaruh tanpa mengambil kira sama ada ianya cerapan berpengaruh baik atau buruk. Tambahan pula, walaupun GDFF boleh mengesan cerapan berpengaruh berganda, ia cenderung mengesan bilangan IO lebih rendah daripada yang sepatutnya disebabkan kesan pengaruh litupan dan limpahan. Kaedah baru yang dicadangkan dan dinamakan Pengubahsuaian Teritlak DFFITS (MGDFF) dibangunkan, yang mana suspek HLPs dalam subset awal dikenalpasti menggunakan kaedah IDRGP yang kami cadangkan.

Sepanjang pengetahuan kami, belum ada kajian dibuat ke atas pengelasan kepada cerapan berpengaruh biasa, baik dan buruk. Oleh itu, plot IDRGP-MGDFF diformulasi bagi menampung lompang dalam literasi.

Tesis ini juga mengenengahkan isu masalah multikolinearan dalam model regresi linear berganda yang berkaitan dengan dua punca. Pertama, ia berpunca disebabkan titik tuasan tinggi, dan punca kedua disebabkan kaedah yang digunakan bagi pengumpulan data, kekangan ke atas model atau dalam populasi, spesifikasi model dan model lampau tertakrif. Walaupun begitu, belum ada kajian memfokus kepada kaedah penganggaran parameter bagi memulihkan masalah multikolineran disebabkan oleh titik tuasan tinggi berganda. Oleh yang demikian, kami mencadangkan kaedah penganggaran baru yang dinamakan Penganggar Terubahsuai GM (MGM) berdasarkan MGDFF. Keputusan kajian menunjukkan penganggar MGM adalah penganggar paling cekap dalam memperbaiki masalah multikolinearan berpunca dari titik tuasan tinggi.

Apabila multikolineran disebabkan oleh punca lain, beberapa kaedah klasikal kedapatan. Di antaranya regresi Ridge (RR), Regresi Jackknife Ridge (JRR) dan regresi Latent Root diketengahkan untuk memulihkan masalah tersebut. Walaubagaimanapun, jelas terbukti bahawa prestasi kaedah penganggaran klasik tersebut sangat lemah dengan kehadiran titik terpencil dalam data. Dalam hal ini, kami mencadangkan dua kaedah penganggaran teguh. Jenis pertama, adalah versi pembaikan LRR bagi memulihkan kedua-dua masalah multikolinearan dan titik terpencil. Kaedah yang dicadangkan diformulasikan dengan menggabungkan penganggar teguh MM dan Pengubahsuaian Penganggar Teritlak M (MGM) dalam algorithma LRR. Kami namakan kedua kaedah ini sebagai kaedah Latent Root berasaskan MM (LRMMB) dan kaedah Latent Root berasaskan MGM (LRMGMB).

Seperti mana jenis pertama, kaedah penganggaran teguh multikolinearan jenis kedua juga berhasrat untuk pembaikan prestasi regresi teguh Jackknife Ridge. Penganggar MM dan penganggar MGM digabungkan dalam algoritma JRR bagi membangunkan versi pembaikian JRR. Kaedah yang dicadangan dinamakan Jackknife Ridge berasaskan MM (JRMMB) dan Jackknife Ridge berasaskan MGM (JRMGMB). Semua kaedah yang telah dicadangkan mengatasi kaedah biasa bila mana multikolineran hadir bersama titik tuasan tinggi berganda.

Pengukuran diagnostik multikolinearan klasik tidak sesuai untuk mendiagnos kewujudan bersama multikolinearan dengan kehadiran titik tuasan tinggi berganda. Apabila VIF klasik digunakan, titik tuasan tinggi menyebabkan peningkatan dan penurunan corak multikolineran. Ini akan mengelirukan kesimpulan dan memberi penunjuk tidak benar dalam menyelesaikan masalah multikolinearan. Maka, kami mencadangkan VIF teguh dinamakan RVIF(JACK-MGM) yang menjadi penunjuk terbaik yang boleh membantu pengamal statistik dalam memilih penganggar yang bersesuaian bagi menyelesaikan masalah multikolinearan.

ACKNOWLEDGEMENTS

First and foremost, I would like to give thanks to my God, who have provided me His strength and grace to throughout my doctoral pursue.

Heartfelt appreciation also goes to my committee chairperson, Prof. Dr. Habshah Midi for her constant inspiration, efficient guidance, and constructive feedback rendered. I am deeply honored to have the opportunity to complete my degree under her supervision.

I would also like to thank my internal co-supervisors, Assoc. Prof. Dr. Isthrinayagy S. Krishnarajah and Dr. Md. Sohel Rana for all their supporting and guidance provided.

My special thanks go to my beloved wife, for standing by with me patiently with her never ending encouragement, prayers and support throughout my doctoral pursue.

Also, I would like to thank my children; Ethar, Kawther and Taha for giving me the happiness during my study.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Habshah bt Midi, PhD
Professor
Facuity of Science
Universiti Putra Malaysia
(Chairman)

Isthrinayagy a/p S. Krishnarajah, PhD
Associate Professor
Facuity of Science
Universiti Putra Malaysia
(Member)

MD. Sohel Rana, PhD
Senior Lecturer
Facuity of Science
Universiti Putra Malaysia
(Member)

BUJANG BIN KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- This thesis is my original work;
- Quotations, illustrations and citations have been duly referenced;
- This thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- Intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- Written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before the thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- There is no plagiarism or data falsification/ fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism software.

Signature: _____ Date: _____

Name and Matric No.: Mohammed Abdulhussein Mohammed GS35192

Declaration by Members of Supervisory Committee

This is to confirm that:

- The research conducted and the writing of this thesis was under our supervision;
- Supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: _____
Name of
Chairman of
Supervisory Committee: Habshah bt Midi, PhD

Signature: _____
Name of
Member of
Supervisory Committee: Isthrinayagy a/p S. Krishnarajah, PhD

Signature:
Name of
Member of
Supervisory Committee: MD. Sohel Rana, PhD

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

BIF	Bounded Influence Function
BIO	Bad Influential Observation
BLP	Bad Leverage Points
BLUE	Best Linear Unbiased Estimators
BP	Breakdown Point
CD	Cook Distance
CI	Confidence Interval
DFFITS	Differences In Fitted values
DRGP	Diagnostic Robust Generalized Potential
GDFF	Generalized DFFITS
GLP	Good Leverage Points
GP	Generalized Potentials
HLPs	High Leverage Estimation
IDRGP	Improvised Diagnostic Robust Generalized Potential
iid	Independent Identically Distributed
IO	Influential Observation
IRLS	Iteratively Reweighted Least Squares
IWLS	Iterative Weighted Least Squares
JRMGMB	Jackknife Robust MGM-estimator Based
JRMMB	Jackknife Robust MM-estimator Based
JRR	Jackknife Ridge Regression
LAD	Least Absolute Deviations
LAR	Least Absolute Residuals
LAV	Least Absolute Values
LMS	Least Median of Squares
LRMGMB	Latent Root MGM-estimator Based
LRMMB	Latent Root MM-estimator Based
LRR	Latent Root Regression
LTS	Least Trimmed Squares
MAD	Median Absolute Deviation
MADN	Normalized Median Absolute Deviation
MCD	Minimum Covariance Determinant
MCD	Minimum covariance Determination
MD	Mahalanobis Distance

MGDFF	Modified Generalized DFFITS
MGM	Modified GM-estimator
MLE	Maximum Likelihood Estimation
MLR	Multiple Linear Regression
MSE	Mean Square Errors
MSE	Mean Square Error
MSGME	Multi-Stage GM-estimators
MVE	Minimum Volume Ellipsoid
MVUE	Minimum Variance Unbiased Estimator
OLS	Ordinary Least Squares
PCE	Percentage of Change in Estimator
RLRR	Latent Root Regression
RLS	Reweighted Least Squares
RMD	Robust Mahalanobis Distance
RO	Regular Observation
RR	Ridge Regression
RRR	Robust Ridge Regression
RVIF	Robust Variance Inflation Factor
S1S	Schweppe one-Step estimator
SD	Standard Deviation
SSE	Sum of Squares Errors
SSR	Sum of Squares Regression
VIF	Variance Inflation Factor
VO	Vertical Outliers
WLS	Weighted Least Squares

CHAPTER 1

INTRODUCTION

1.1 Introduction and Background of the Study

Regression analysis is basically a statistical technique for investigating the functional relationship among two or more quantitative variables so that, a dependent or response variable can be predicted from one or more of predictor or explanatory variables (Kutner et al., 2005), where the predictor variables assumed to be fixed. Regression analysis involves model building, parameter estimation and prediction. The ordinary least squares (OLS) is one of the predominant regression analysis techniques. When the Gaussian Markov assumptions are met, the OLS is the most popular estimation method in linear regression model due to its supreme properties and ease of computation. In addition, when the random errors are independent identically distributed (*iid*) normal, the OLS estimator is WKH EHVW OLQHDU XQELDVHG HVWLPDWRUV (BLUE). In other words, the OLS estimator has the smallest variance among all possible linear unbiased estimators. Furthermore, the maximum likelihood estimator (MLE) equals the OLS estimator under these conditions. Unfortunately, in real practice the assumptions about the normality of the error term distribution and the independency of the predictor variables (multicollinearity problem) are always violated. Furthermore, the OLS estimator is not robust against unusual data and it has very low breakdown point which is equals to 1/*n* (Maronna, 1976), where *n* is the size of the sample.

Even one unusual observation can drastically change the OLS estimate very badly (see Rousseeow and Leroy, 1987; Gujarati, 2003; Kamruzzaman and Imon, 2002; Kutner et al., 2005; Maronna et al., 2006; Andersen, 2008).

The assumption of normality is violated in the presence of one or more influential observations. Belsley et al. (1980) stated that influential observations were those observations either alone or together with several other observations have the largest impact on the computed values of various estimates. Barnett and Lewis (1994) defined outliers as those observations that are markedly far from the majority of observations in a data set.

There are several versions of outliers in regression problems. Observations are judged as residual outliers based on how unsuccessful the fitted regression equation is in accommodating them. This is the reason why observations corresponding to very large residuals are treated as residual outliers. Observations which are extreme or outlying in the *y*-coordinate are called outliers or vertical outliers. High leverage points (HLPs) are those observations which are outlying in the *X*-coordinate. It is often very essential in regression analysis to find out whether HLPs; those observations which fall far from the majority of the independent variables, have much impact on the fitting of a model. HLPs not only fall far from the majority of predictor variables, but also are deviated from a regression line (Belsley et al., 1980; Hocking and Pendelton 1983; Rousseeow and Leroy, 1987).

The other serious problem is that HLPs have high impact on the OLS estimators and is responsible for causing multicollinearity problem. Multicollinearity is a situation of multiple regression model when the independent variables are highly correlated. If the purpose of a study is to predict response variable (y) from a group of explanatory variables (*x*), the multicollinearity is not problematic. Nevertheless, if the purpose is to illustrate the impact of individual *x* variable on y, then the multicollinearity is a big problem. Imon and Khan (2003) exemplified that HLPs is a new source of multicollinearity. These leverage points may increase (enhancing observation) or decrease (reducing observation) multicollinearity problem (Habshah et al., 2011).

In real-life situations, multicollinearity, the existence of anomalous points and departure from the normality assumption are common problems in regression analysis. This fact is pointed out by many standard books, articles and researchers. Nowadays, several procedures which deal with multicollinearity and outliers separately are available. However, there is not much significant work reported in the literature which takes into account the presence of both multicollinearity and outlier problems simultaneously (see Johnston, 1984; Montgomery et al., 2001; Gujarati, 2002; Kutner et al., 2005; Chatterjee and Hadi, 2006; Kamruzzaman and Imon, 2002, Imon, 2005).

1.2 Importance and Motivation of the Study

Linear regression analysis is the most significant statistical technique in many fields such as economics, survival studies, business, medicine, engineering and others. To estimate the coefficients of the linear regression model, the least squares method is often used because of tradition and it is easy to compute. However, in the presence of single or multiple enormous points in a data set can destroy the OLS estimates. Many researchers stated that a real data set usually contain 1% to 10% of unusual observations (Hampel et al., 1986; Wilcox, 2005). HLPs have more serious effects on the OLS estimates than the outliers in y variable. According to Pena and Yohai (1995), HLPs are responsible for masking and swamping of outliers in linear regression. HLPs are also causing multicollinearity problem and have great effect on the values of various estimates. Hence, it is vital to detect those unusual observations. Although, ±DGL¶Ìotential values (Hadi, 1992) can detect single leverage point but they are not successful to identify multiple HLPs due to masking and swamping effects (Rousseeuw and Leroy, 1987; Ruppert and Simpson, 1990; Imon, 1996; Imon, 2005, Habshah et al. 2009). To address this problem, Imon (1996) suggested the generalized potentials (GP) as a diagnostic method for detecting multiple HLPs. The idea of generalized potential is by extending a single case deletion to a group of case deletion. Habshah et al. (2009) pointed out that the GP approach is not very successful in identifying the correct number of HLP due to its inefficient way of choosing the initial basic subset, which suffers from masking effects. To remedy this problem, Habshah et al. (2009) developed the diagnostic robust generalized potential (DRGP) approach which is very successful in identifying HLPs. Nonetheless, the DRGP approach has small rate of masking and swamping effects, especially with small size of sample and high percentage of contamination. This shortcoming of DRGP has

inspired us to develop a new technique to improve the performance of DRGP which we call the improvised DRGP (IDRGP). It is proposed by adding new step in the DRGP algorithm pertaining to two cases. The proposed IDRGP is expected to show higher rate of detection of HLPs with smaller masking and swamping rates.

This thesis also concerned on the diagnostic measures for the identification of bad influential observations (BIO). The detection of BIO is very important because it is accountable for inaccurate prediction and invalid inferential statements as it has large impact on the computed values of various estimates. The Generalized version of DFFITS (GDFF) which is proposed by Imon (2005) is developed only to identify influential observations (IO) without taking into consideration whether it is good or bad. In addition, although GDFF can detect multiple IO, it has a tendency to detect lesser IO as it should be. This is due to the choice of the initial basic subset of the GDFF which is not adequately effective in classifying the deletion and the remaining groups. The weakness of GDFF is anticipated to improve the detection rate of IO while keeping smaller masking and swamping effects. The new proposed method is called modified generalized DFFITS (MGDFF), whereby the suspected HLPs (HLP) in the initial subset are identified using our proposed IDRGP diagnostic method.

Statistics practitioners are often rely on handy plot to quickly capture irregularities in a data set. A diagnostic plot is very useful in this regard. Russeeuw and Van Zomeren (1990) proposed the LMS-RMD plot to classify observations into regular observations, vertical outliers, good and bad high leverage points. Nonetheless, this plot is based on the RMD which is known to suffer from masking and swamping effects. Bagheri and Habshah (2015) suggested the LTS-DRGP(MVE) plot for identifying such points. This plot depends on the DRGP which has small swamping effects. Moreover, both plots do not specifically constructed to classify observations into regular, vertical, good and bad influential observations. To the best of our knowledge, no such plot is found in the literature. The limitation of these plots and literature gap has encouraged us to formulate a new classification scheme which we call IDRGP-MGDFF to classify observations into good and bad influential observations. The performance of the MGDFF is assessed through real data and simulation study.

This thesis also addresses the issue of multicollinearity problem in multiple linear regression models. The OLS estimator suffers a huge set back in the presence of multicollinearity. There are many sources of multicollinearity problem, such as the data collection method employed, constraints on the model or in the population, model specification and an over defined model (Montgomery et al., 2001). However, there is an evident that the high leverage point is another source for multicollinearity (Imon and Khan, 2003). Irrespective of the source, when multicollinearity problem is detected it is obvious to remedy this problem in order to obtain efficient parameter estimates. Imon and Khan (2003) proposed using generalized potentials (GP) measure to overcome the multicollinearity problem caused by the presence of multiple HLPs. As already mentioned the drawback of this measure is that it is not very successful in identifying correct HLPs and suffers from masking and swamping effects. To

the best of our knowledge, there do not exist any literature in multicollinearity which discusses solution to the problem of multicollinearity caused by HLPs. This inspires us to propose a new estimation method namely the modified GM-estimator (MGM) based on MGDFF to combat the multicollinearity caused by the multiple HLPs. When multicollinearity is due to other sources not HLPs, several classical methods are available. Among them, the Ridge Regression (RR), Jackknife Ridge Regression (JRR) and Latent Root Regression (LRR) are put forward to remedy this problem. Nevertheless, it is now evident that these classical estimation methods perform poorly when outliers exist in a data. Not much work is devoted when multicollinearity comes together with the existence of outliers. In this situation, we propose two types of robust estimation methods. The first type is an improved version of the LRR to rectify the simultaneous problems of multicollinearity and outliers. The proposed method is formulated by incorporating robust MM-estimator and the modified generalized M-estimator (MGM) in the LRR algorithm. We call these methods the Latent Root MM-based (LRMMB) and the Latent Root MGM-based (LRMGMB) methods. Similar to the first type, the second type of robust multicollinearity estimation method also aim to improve the performance of the robust jackknife ridge regression. The MM-estimator and the MGM-estimator are integrated in the JRR algorithm for the establishment of the improved versions of JRR. The suggested method is called jackknife ridge MM based denoted by JRMMB and the jackknife ridge MGM based denoted by JRMGMB.

The classical multicollinearity diagnostic methods may not be suitable to correctly diagnose the existence of multicollinearity in the presence of HLPs (Montogmery and Askin, 1981; Rosen, 1999). When we use the classical multicollinearity diagnostic methods, the HLPs may increase (high leverage collinearity-enhancing observation) or decrease (high leverage collinearity-reducing observation) the multicollinearity pattern of a data. Subsequently, the classical VIF gives incorrect indicator for solving multicollinearity problem because statistics practitioners often rely on this diagnostic measure. To the best of our knowledge, not much work is devoted to robust VIF. Bagheri and Habshah (2011) proposed RVIF(MM) and RVIF(GM(DRGP)) to diagnose multicollinearity. Nonetheless, the RVIF(MM) is not efficient (Bagheri and Habshah, 2011). The RVIF(GM(DRGP)) is expected not to be very efficient because it is formulated based on DRGP that has been proven in Chapter 3, less efficient than IDRGP. Moreover, it is also based on GM(DRGP) estimator which is less efficient as it downweight all detected HLPs irrespective of whether it is good or bad. This issue has inspired us to develop a new robust VIF, namely the RVIF(jack-MGM) which is anticipated to be more reliable than the RVIF(GM(DRGP) because it is based on Jack-MGM estimator which is proven in Chapter 5 to do credible job.

1.3 Research Objectives

The main aim of this thesis is to investigate the multicollinearity problems for linear regression model in the presence of HLPs. The classical diagnostic and estimation methods which deal with multicollinearity problems are mostly based on ordinary least squares (OLS) estimates. Unfortunately, the OLS estimate is not robust for HLPs. Moreover, there is evidence that HLPs is a new source for

multicollinearity. It is important to modify the classical diagnostic multicollinearity methods to be more resistance for HLPs. In addition, it will be interesting to develop a new technique to detect the correct number of HLPs in a data set and to classify correctly the HLPs into good and bad influential observations. The foremost objectives of our research can be outlined systematically as follows:

1. To propose a new improved diagnostic measure for the identification of multiple HLPs in order to obtain the exact number of multiple HLPs that able to reduce masking and swamping effects.
2. To formulate a new diagnostic measure to identify multiple influential observations.
3. To propose a new classification scheme to classify observations into 4 types: regular observation, vertical outliers, good and bad influential observation.
4. To propose a new robust estimation method to solve the multicollinearity problem that is due to HLPs.
5. To formulate a robust latent root regression and robust jackknife ridge regression estimation techniques for linear regression having both multicollinearity and HLPs.
6. To develop robust multicollinearity diagnostic measures for detecting multicollinearity problem in the presence of HLPs.

1.4 Scope and limitation of the study

The multiple linear regression model is widely used in many fields of studies such as business, economics, medicine and social sciences. In real situation, it has many practical uses. However, the most application is to fit the predictive model to an observed data set of response and predictor variables. Multiple linear regressions are predominantly fitted using the OLS method because of tradition and ease of computation. When the underlying assumptions are hold, the OLS estimates have the optimal properties. In reality, the underlying assumptions of independency among the predictor variables and normality of the random errors are always violated. In addition, the OLS estimate is not resistant to outlying observations. Even one outlier can make a big changed in the OLS estimate. As an alternative methods, many robust statistical estimation techniques are suggested such as, least median of squares, least trimmed squares, S-estimator, M-estimator and MM-estimator. Nonetheless, most of the existing methods alone cannot be used to remedy the combined problem of outliers in the presence of multicollinearity. On the other hand, there are good numbers of work on the identification of HLPs (Ellenberg, 1976; Belsley et al. 1980; Rousseeuw, 1987; Imon, 2002; Imon 2005; Habshah et al. 2009). Nevertheless, most of these detection methods basically focused only on the identification of HLPs without taking into consideration their classification into good and bad leverage points. It is very important to detect and classify the good and bad leverage points, as only bad leverage points are responsible for the misleading conclusion about the fitting of the regression model.

Since robust statistic is relatively new technique in statistics, there are not so many algorithms and statistical softwares which are available for complicated robust applications. In addition, not many outlying data sets are available in the literatures. Furthermore, only few outlying data sets with multicollinearity problems are available. For these reasons, the same data sets were used repeatedly for different objectives of this study.

1.5 Overview of the Thesis

In accordance with the objectives and the scope of the study, the contents of this thesis are structured in the eight chapters. The thesis chapters are organized so that the study objectives are apparent and are conducted in the sequence outline.

Chapter Two: This chapter briefly presents the literature review of the least squares estimation method and the violations from its underlying assumptions such as departure of normality and presence of multicollinearity problem. The outliers, HLPs, influential observation and their diagnostics methods are also discussed. Moreover, basic concepts of robust regression and some important existing robust regression methods are also reviewed. The effects, sources and consequences of multicollinearity and its estimation and diagnostics methods are also highlighted. Finally, bootstrapping methods are also briefly discussed.

Chapter Three: This chapter discusses the existing DRGP which is developed by Habshah et al. (2009). The new proposed method, the improvised DRGP (IDRGP), for the identification of multiple HLPs is presented. The steps for IDRGP and its algorithm are also highlighted. Finally, some examples artificial data and A Monte Carlo simulation study are discussed.

Chapter Four: In this chapter**,** a modified generalized DIFFITS (MGDFF) based on IDRGP for the identification of good and bad influential observations is developed. A new IDRGP-MGDFF Scheme of classifying observations into regular observations, vertical outlying observations, good and bad influential observations is also presented. Some real data, artificial data and A Monte Carlo simulation study are discussed to assess the performance of our proposed method.

Chapter Five: This chapter deals with the development of the GM-estimator based on modified generalized DIFFITS (denoted by MGM) for data having multicollinearity due to HLPs. A Monte Carlo simulation study and two numerical examples are carried out to assess the performance of the proposed method.

Chapter Six: In this chapter, two types of robust multicollinearity estimation methods are formulated. The first type deals with the development of robust latent root regression estimation methods; LRMMB and LRMGMB. This method is formulated by incorporating the high efficient and high breakdown MM-estimator and our proposed MGM-estimator methods in the classical latent root regression, respectively. The second type deal with the robust jackknife ridge regression estimation methods. In this respect we propose two robust

methods namely; JRMM and JRMGM. These robust methods are formulated by integrating the classical jackknife ridge regression method with the MM-estimator and MGM-estimator, respectively. A Monte Carlo simulation study and some numerical examples are given to assess the performance of the proposed method.

Chapter Seven: In this chapter, we present the proposed multicollinearity diagnostic measures, namely the robust VIF based on robust jackknife ridge regression (denoted by RVIF(Jack-MGM)).The new proposed measures are useful to detect the multicollinearity problem in the presence of influential observations. In this respect, two types of data are considered, the collinearity data with reducing influential observations and non-collinearity data with enhancing influential observations. The numerical results and Monte Carlo simulation are also discussed to assess our proposed measures.

Chapter Eight: This chapter provides summary and detailed discussions of the thesis conclusions. Areas for future research are also recommended.

REFERENCES

Aelst, S. V. and Rousseeuw, P. (2009). Minimum volume ellipsoid. *Advanced Review*. Vol. 1: July/August 2009.

Anderson, C. (2001). *A Comparison of Five Robust Regression Methods with Ordinary Least Squares: Relative Efficiency, Bias, and Test of the Null Hypothesis*, Unpublished Ph.D. thesis, University of North Texas, U.S.A.

Andersen, R. (2008). *Modern methods for robust regression*. The United States of America: Sara Miller McCune. SAGE publications.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics*.16:523-531.

Andrews, D. F., & Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society. Series B (Methodological)*, 85-93.

Askin, R. G. and Montgomery, D. C. (1980). Augmented robust estimators, *Technometrics.* 22: 333-341.

Atkinson, A. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, 89(428), 1329-1339.

Baghri, A. (2011), *Rbust Estimation Methods and Robust Multicollinearity Diagnostics For Multiple Regression Model in The Presence of High Leverage Collinearity -Influential Observation*, Thesis submitted to the School of Graduate Studies, UPM.

Bagheri, A., Habshah M., & Imon, R. H. M. R. (2012). A novel collinearity-influential observation diagnostic measure based on a group deletion approach. *Communications in Statistics-Simulation and Computation*, 41(8), 1379-1396.

Bagheri, A. and Habshah, M. (2015). Diagnostic plot for the identification of high leverage collinearity- influential observation, *Statistics and Operations Research Transactions*, 39(1): 51-70.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. 3rd edition. New York: Wiley.

Bastlevsky A. (1994), *Statistical Factor Analysis And Related Method*s, John Wilay & Sons, INC.

Batah F. S., Ramanathan T. V. and Gore S. D. (2008). The e¥ ciency of PRGLHGMDFNQLHDQGULGJHWSHUHJUHVVLRQHVWLPDWRUVDFRPSDULV Surv. Math. Appl, 3:111-122.

Beaton, A.E. and Tukey, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*.16: 147-185.

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

Belsley, D.A., Kuh, E. and Welsch, R.E. (2004). *Regression Diagnostics, Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, New York.

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

Birch, J. B. (1980). Some convergence properties of iterated reweighted least squares in the location model. *Communications in Statistics - Simulation and Computers*. B9: 359-369.

Birch, J. B. and Agard, D. B. (1993). Robust inference in regression: a comparative study. *Communications in Statistics- Simulation and Computers*. 22: 217-244.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*. 40: 318-35.

Brown, P. J. (1977). Centering and scaling in ridge regression. *Technometrics*.19:35-36.

Butler, R.W., Davies, P.L. and Jhun, M. (1993). Asymptotics for the Minimum Covariance Determinant estimator. *Annals of Statistics*. 21:1385±1400.

Campbell, N. A. (1980). Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation. *Applied Statistics*. 29: 231±237.

Campbell, N. A., Lopuhaä, H. P. and Rousseeuw, P. J. (1998) . On the calculation of a robust S-estimator of a covariance matrix. *Statistics in Medicine*. 17(23): 2685-2695.

Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression*. NewYork:Wiley.

Chatterjee, S. and Hadi, A.S. (2006). *Regression Analysis by Example*. 4th edition. New York: Wiley.

Coakley, C. W. and Hettmansperger, T. P. (1993). A bounded-influence, high-breakdown, efficient regression estimator. *Journal of the American Statistical Association*. 88:872±880.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*. 19:15-18.

Cook, R. D. **(**1979). Influential observations in linear regression. *Journal of American Statistical Association*. 74:169-174.

Cook, R. D. and Hawkins, D. M. (1990). Unmasking multivariate outliers and leverage points: Comment. *Journal of the American Statistical Association*. 85: 640-44.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Champan Hall.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the Minimum Covariance Determinant scatter matrix estimator. *Journal of Multivariate Analysis*. 71:161±190.

Daszykowski, M. , Kaczmarek, K., Vander, Y. H. and Walczak, B. (2007). Robust statistics in data analysis - a review basic concepts. *Chemometrics Intelligence Labo*ratory. 85: 203-219.

Davies, P. L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Annals of Statistics.* 15:1269±1292.

Donoho, D. L. (1982). *Breakdown Properties of Multivariate Location Estimators*. Unpublished Ph.D. thesis, Harvard University, The American United States.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. New York:Wiley.

Edgeworth, F. Y. (1887). On observations relating to several quantities. *Hermathena*. 6:279-285.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*.7:1-26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canadian Journal of Statistics*. 9:139-72.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association.* 82: 171±185.

Efron, F. (1982). *The jackknife. the Bootstrap, and other resampling plans*. Philadelphia. Penn.: Society for industrial and applied mathematics.

Ellenberg, J.H. (1976). Testing for a single outlier from a general linear regression. *Biometrics*. 32: 637-645.

Fox, J., & Weisberg, S. (2010). *An R companion to applied regression*. Sage.

Gray, J. B. (1985). Graphics for regression diagnostics. *Proceedings of Statistical Computing Section*, pp. 102-107.

Linear Regression- Lecturer Notes in Statistics, Springer Verlag Berlin Heidelberg.

Greene, W. H. (2008). *Econometric analysis*. 6th edition. Upper saddle river. New jersey: Prentice Hall.

Grewal, R., Cote, J.A. and Baumgartner, H. (2004). Multicollinearity and measurement error in structural equation models: Implications for Theory Testing . *Marketing Science*. 23(4): 519-529.

Gross, J. (2003). *Linear regression*. 1th edition. New York: Springer-Verlag.

Gujarati, D.N. (2002). *Basic Econometrics*. 4th edition. New York: Macgraw-Hill.

Gunst, R. F., Webster, J. T. and Mason, R. L (1976), A comparison of least squares and latent root regression estimators, *Technometrics*, l(18): 75-83.

Gunst, R.F. (1983). Regression analysis with multicollinear predictor variables: definition, detection,and effects. *Communications in Statistics: Theory and Methods*. 12:2217±2260.

Habshah M., Bagheri, A., & Imon, A. H. M. (2010). The Application of Robust Multicollinearity Diagnostic Method Based on Robust Coefficient Determination to a Non-Collinear Data. *Journal of Applied Sciences*, 10(8):611-619.

Habshah M. and Marina Z. (2007). $6LPXODWLRQVWXGLHVUHJUHVVLRQ HVWLPDWLRQLQWKHSUHVHQFHRIRXWOLHUVDQGPXOWLFROOLQHDULW *Teknologi*, 47: 59-74.

Habshah, M., Norazan, M.R. and Imon, A.H.M.R. (2009). The performance of Diagnostic-Robust Generalized Potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics.* 36(5): 507-520.

Hadi, A.S. (1988). Diagnosing collineariy-influential observations. *Computational Statistics and Data Analysis*. 7:143-159.

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

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*. 69: 383-393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*. New York: Wiley.

Handshin, E., Schweppe, F. C., Kohlas, J. and Fiechter, A. (1975). Bad data analysis for power system state estimation. *EEE Transactions of Power Apparatus and Systems.* PAS-94: 329-337.

Hocking, R.R. & Pendelton, O.J. (1983). The regression dilemma,Communications in Statistics-*Theory and Methods* 12: 497-527.

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.

Hill, R. W. (1977). *Robust Regression When There Are Outliers in the Carriers*. Unpublished Ph.D. thesis. Harvard University, Boston, MA.

Hill, R. W. and Holland, P. W. (1977). Two robust alternatives to robust regression. *Journal of the American Statistical Association.* 72: 828± 833.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics*, 19(3), 285-292.

Hoaglin, D. C and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *American Statistician.* 32:17-22.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: applications to non-orthogonal problems. *Technometrics*. 12:69-82.

Hoerl, A. E., and Kennard, R. W. (1970b). Ridge regression: biased estimation for non- orthogonal problems. *Technometrics*.12:55-67.

Hogg, R. V. (1979). *An introduction to robust estimation. Robustness in Statistics.* eds. R. Launer and G. Wilkinson. Academic Press: New York.

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9), 813-827.

Huber, P. J. (1964). Robust estimation of location parameters. *Annals of Mathematical Statistics.* 35:73±101.

Huber, P. J. (1973). Robust regression: asymptotic, conjectures, and Monte Carlo. *The Annals of Statistics*, 1: 799-821.

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

Huber, P.J (2003). *Robust Statistics,* Wiley, New York, USA.

Huber, P.J. (2004). *Robust Statistics*. New York: John Wiley & Sons.

Imon, A.H.M.R. (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies*. *Special Volume in Honour of Professor Mir Masoom Ali*. 3: 207±218.

Imon, A.H.M.R. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied Statistics*. 32: 929-946.

Imon, A.H.M.R. and Khan, M.A.I. (2003). A solution to the problem of multcollinearity caused by the presence of multiple high leverage points. *International Journal of Statistical Science*. 2: 37-50.

Jadhav N. H. and Kashid D. N. (2011). A Jackknifed Ridge M-Estimator for Regression Model with Multicollinearity and Outliers, *Journal of Statistical Theory and Practice*,(5): 659-673.

Johnston, J. (1984). *Econometric Methods.* 3rd Edn.. New York: McGraw Hill.

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

Katz, M. H. (2006). *Multivariate Analysis: a Practical Guide for Clinicians*. UK: Cambridge University Press.

Kempthorne, P. J. and Mendel, M. B. (1990). Unmasking multivariate outliers and leverage points: Comment. *Journal of the American Statistical Association*. 85: 647-48.

Kent, J. T. and Tyler, D. E. (1996). Constrained M-estimation for multivariate location and scatter. *Annals of Statistics*. 24:1346±1370.

Kim, M. G. (2004). Sources of high leverage in linear regression model. *Journal of Applied Mathematics and Computing*.16(1- 2): 509-513.

Krasker, W. S. and Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *Journal of the American Statistical.* 77: 595± 604.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005). *Applied Linear Regression Models*. 5th edition. New York: MacGRAW-Hill.

Lawrence K. D. and Arthur, J. L. (1990). *Robust Regression; Analysis and Applications*, INC: Marcel Dekker.

Lawrence, K. D. and Marsh, L. C. (1984). Robust ridge estimation methods for predicting U. S. coal mining fatalities. *Communications in Statistics-Theory and Methods.* 13: 139-149.

Li, G., and Chen, Z. (1985). Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo, *Journal of the American Statistical Association*. 80: 759±766.

Lupuhaä, H. P. (1991). 2Œ Estimators for location and scatter. *Canadian Journal of Statistics.* 19:307-321.

Lupuhaä, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *Annals of Statistics.* 27:1638-1665.

Mallows, C. L. (1975). *On Some Topics in Robustness*. Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

Maronna, R. A. (1976).Robust M-Estimators of Multivariate Location and Scatter. *The Annals of Statistics*. 4: 51±67.

Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics Theory and Methods*. New York: Willy and sons.

Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high- dimensional datasets. *Technometrics.* 44:307±317.

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*.12: 591-612.

Mason, R.L. and Gunst, R.F.(1985). Outlier-induced collinearities. *Technometrics*. 27:401±407.

McDonald G. C[illegible] some ridge-[illegible], 20: 407-416.

Miller, R. G. (1974). The jackknife-a review. *Biometrika*, 61(1), 1-15.

Moller, S.F., Frese, J.V. and Bro, R. (2005). Robust Methods for multivariate data analysis. *Journal of Chemometrics*. 19(10): 549-563.

Montgomery, D. C and Askin, R. G. (1981). Problems of nonnormality and multicollinearity for forecasting methods based on least squares. *AIIE Transactions.* 13:102-115.

Montgomery, E., Bronner, M. P., Goldblum, J. R., Greenson, J. K., Haber, M. M., Hart, J. & Washington, K. (2001). *Reproducibility of the diagnosis of dysplasia in Barrett esophagus: a reaffirmation.* Human pathology, 32(4), 368-378.

Montgomery, D. C., Peck, E. A. and Viving, G.G. (2001). *Introduction to linear regression Analysis*. 3rd edition. New York: John Wiley and sons.

Mosteller, F. and Tukey , J. W. (1977) . *Data Analysis and Regression*. Reading. MA: Addison- Wesley Publishing Company.

Myers, R. H. (1990). *Classical and Modern Regression with Applications*. 2nd edition. CA: Duxbury press.

Meyers, L. S., Gamst, G. and Guarino, A. J. (2006). *Applied Multivariate Research: Design and Interpretation*. Sage publications, INC.

Neter, J., Kutner, M.H., Wasserman W. and Nachtsheim, C.J. (2004), *Applied Linear Regression Models*. New York: MacGRAW-Hill/Irwin.

Odutan,G.(2004). *A Monte Carlo Study of the Problem of Multicollinearity in a Simultaneous Equation system.* Unpublished Ph.D., Thesis. University of Ibadan, Nigeria.

Olive, D. J.(2008). *Applied Robust Statistics*, Southern Illinois University. http://www.math.siu.edu/olive/run.pdf

Pearson, K. (1908). On the generalized probable error in multiple normal correlation. *Biometrika*. 6: 59-68.

Pearson, E. S. **(**1931). The analysis of variance in cases of non-normal variation. *Biometrika*. 23:114-133.

Peña, D. and Yohai, V.J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of Royal Statistical Society*. B 57: 18±44.

Pfaffenberger, R. C, and Dielman, (1985) T. E. , A Comparison of Robust Ridge Estimators, *Business economics section Proceedings of the American Statistical Association.* 631 -635.

Rao, C. R. (1965). *Linear statistical inference and its applications*, John Wiley & Sons.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 353-360.

Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, *91*(435),1047-1061.

Rosen, D.H. (1999). *The Diagnosis of Collinearity: A Monte Carlo Simulation Study*, Department of Epidemiology, Unpublished Ph.D. thesis, School of Emory University. Atlanta, USA.

Pison, G., Rousseeuw, P. J., Filzmoser, P., & Croux, C. (2003). Robust factor analysis, *Journal of Multivariate Analysis*, 84(1), 145-172.

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

Rousseeuw, P. J. (1983). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*. Vol (B): 283-297.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of theAmerican Statistical Association.* 79: 871±880.

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. *Mathematical and Statistical Applications*. B: 283-297.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute values. *Journal of the American Statistical Association*. 88: 1273-83.

Rousseeuw P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics.* 41:212±223.

Rousseeuw, P. and Van Zomeren, B. (1990).Unmasking multivariate outliers and leverage points. *Journal of American Statistical Associations*. 85: 633-639.

Rousseeuw, P. J. and Yohai, V. (1984). Robust regression by means of S-estimators, Robust and Nonlinear Time series Analysis. *Lecture Notes in Statistics*. 26: 256-272.

Ruppert, D. and Simpson, D. G. (1990). Unmasking multivariate outliers and leverage points: Comment. *Journal of the American Statistical Association*. 85:644-646.

Ryan, T. P. (1997). *Modern Regression Methods*. NewYork: Wiley.

Ramsay, J.O.(1977). A comparative study of several robust estimates of slope, intercept, and scale in linear regression, *Journal of American Statistical Associations.*72:608-615.

Sall, J. (1990). Leverage plots for general linear hypothesis. *The American Statistician*. 44: 308-315.

Sengupta, D. and Bhimasankaram, P. (1997).On the roles of observations in collineariy in the linear model. *Journal of American Statistical Association*. 92:1024-1032.

Simpson, J. R. (1995). *New Methods and Comparative Evaluations for Robust and Biased-Robust Regression Estimation*. Unpublished Ph.D. thesis, Arizona State University, The United States of America.

Simpson, D. G., Ruppert, D. and Carroll, R. J. (1992). On one-step GM estimates and stability of influences in linear regression. *Journal of the American statistical association*. 87: 439-450.

6LQJK%&KDXEH\3DQG'ZLYHGL7$QDOPRVWXQELDVHGULGJH HVWLPDWRU *The Indian Journal of Statistics*,(48): 342-346.

Srikantan, K.S. (1961). Testing for the single outlier in a regression mod *Sankyā* *Series* A, 23: 251±260.

Stahel, W. A. (1981). *Breakdown of Covariance estimators*, Research report 31, Fachgruppe für Statistik, Swiss Federal Institute of Technology (ETH):Zürich.

Stevens, J. P. (2002). *Applied Multivariate Statistics for the Social Sciences*. 4th edition. Hillsdale .NJ: Erlbaum.

Stine, R. A. (1995). Graphical Interpretation of Variance Inflation Factors. *The American Statistician.* 49: 53-56.

Stromberg, A. J., Hossjer, O. and Hawkins, D. M. **(**2000). The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association*. 95: 853-864.

Tyler, D. E. (1994) . Finite sample breakdown points of projection-based multivariate location and scatter statistics. *Annals of Statistics*. 22:1024-1044.

Vellman, P.F. and Welsch, R.E. (1981). Efficient computing of regression diagnostics. *American Statistician*. 27: 234-242.

Walker, E. (1985). *Influence, Collinearity and Robust Estimation in Regression*. Unpublished Ph.D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Walker, E. (1989). *Detection of collinearity-influential observations*. Communications in Statistics-Theory and Methodology.18:1675-1690.

Wang, N. and Raftery, A. E. (2002). Nearest-neighbor variance estimation (NNVE): Robust covariance estimation via nearest-neighbor cleaning. *Journal of the American Statistical Association.* 97: 994-1006.

Webster, John T., Gunst, Richard F., and Mason (1974), Robert L, Latent Root Regression Analysis, *Technometrics*, 16(4): 513-22.

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*. 2nd edition. The United States of America: Elsevier academic.

Welsch, R. E. (1980). Regression sensitivity analysis and bounded-influence estimation. In *Evaluation of econometric models* (pp. 153-167). Academic Press.

Woodruff, D. L. and Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association.* 89:888±896

Yohai, V.J. (1987). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*. 15: 642-656.