ROBUST STATISTICAL PROCEDURES FOR MULTIPLE LINEAR REGRESSION MODEL IN THE PRESENCE OF MULTICOLLINEARITY AND OUTLIERS

By

HENDI HASAN TALIB HENDI

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

December 2023

IPM 2023 6

DEDICATION

To my parents, brothers and all the persons who prayed for me and supported me during my study journey

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

ROBUST STATISTICAL PROCEDURES FOR MULTIPLE LINEAR REGRESSION MODEL IN THE PRESENCE OF MULTICOLLINEARITY AND OUTLIERS

By

HENDI HASAN TALIB HENDI

December 2023

Chairman : Professor Habshah binti Midi, PhD
Institute : Mathematical Research

In this study, a new version of diagnostic method of identification of multiple high leverage points (HLPs) is developed by incorporating the location and scatter estimators obtained from the $\sqrt{n}$ Reweighted Fast Consistent and High Breakdown method. To improve the accuracy of the identification of HLPs, a robust measure of scale, Q_n is integrated in the computation of cut-off point for the new method. The results indicate that the new method successfully identify high leverage points with the least computing time, highest percentage of correct detection of high leverage points and smallest percentage of swamping and masking effects compared to the existing diagnostic Robust Generalized Potential method which is based on the Minimum Volume Ellipsoid and diagnostic Robust Generalized Potential method which is based on the Index Set Equality.

Several methods have been developed to detect multiple influential observations in linear regression that includes the Fast Improvised Influential Distance. However, this method is computationally not stable, still suffers from masking and swamping effects, time consuming issues and not using proper cut-off point. As a solution to this problem, a new robust version of influential distance method which is based on $\sqrt{n}$ Reweighted Fast Consistent and High Breakdown estimator is proposed. A confidence bound type cut-off point is also proposed. The results signify that the proposed method is very successful in identifying multiple influential observations with the least computational running times, least swamping effect and no masking effect compared to Influential Distance and Fast Improvised Influential Distance methods.

The Generalized M-estimator (GM6) poses several weaknesses such as it down weights good high leverage points and takes very long computational running times. The Generalized M- estimator based on Fast Improvised Generalized Studentized residual

(GM-FIMGT) is introduced to remedy this problem. However, this estimator uses Index Set Equality in the computation of robust Mahalanobis distance which is found to be not stable and still possesses masking and swamping effect. Thus, a new robust GM estimator (GM-RFIID) that incorporates a new weight function constructed from a new version of influential distance method is established. The findings show that the GM-RFIID is more efficient than the GM-FIMGT and GM6 estimators.

Robust Jackknife Ridge MM and Jackknife Ridge GM2 estimators are developed to simultaneously rectify the problem of multicollinearity with the existence of outliers. Nonetheless, both estimators are still not very efficient because they suffer from long computational running time and not very successful in handling high leverage points. Hence, a robust Jackknife Ridge Regression based on GM-RFIID estimator is proposed. The results show that the proposed estimator outperforms the existing methods discussed in this thesis.

Not much research has been done to address the multicollinearity issue caused by high leverage points. The deletion of suspect HLPs from the analysis by using Generalized Potentials, the Generalized M-Diagnostic Robust Generalized Potential-Least Trimmed of Squares and the Diagnostic Robust Generalized Potential-MM methods are the three methods that are found in the literature. However, they are not very efficient. Hence, we propose the GM-RFIID method to remedy the problem of multicollinearity which is caused by high leverage points. The results show that the GM-RFIID method is superior compared to the existing methods discussed in this thesis.

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

PROSEDUR BERSTATISTIK TEGUH BAGI MODEL REGRESI LINEAR BERGANDA DENGAN KEHADIRAN MULTIKOLINEARAN DAN TITIK TERPENCIL

Oleh

HENDI HASAN TALIB HENDI

Disember 2023

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

Dalam kajian ini, kaedah berdiagnostik versi baharu bagi mengesan titik tuasan tinggi berganda dibangunkan dengan menggabungkan penganggar lokasi dan skala yang diperoleh daripada penganggar $\sqrt{n}$ berpemberat pantas konsisten dan kaedah titik musnah tinggi. Bagi meningkatkan ketepatan pengecaman titik tuasan tinggi, ukuran skala teguh, Q_n digabungkan dalam pengiraan titik-potong bagi kaedah baharu ini. Keputusan menunjukkan bahawa kaedah baharu ini berjaya mengesan titik tuasan tinggi dengan masa pengiraan yang paling singkat, peratus tertinggi pengesanan titik tuasan tinggi yang betul dan peratus kesan limpahan dan penyorokan paling kecil berbanding dengan Kaedah Potensi Teritlak Teguh Berdiagnostik yang berasaskan Isipadu Minimum Ellipsoid dan Kaedah Potensi Teritlak Teguh Berdiagnostik yang berasaskan indeks set kesaksamaan.

Beberapa kaedah telah dibangunkan bagi mengesan cerapan berpengaruh berganda dalam regresi linear termasuk Kaedah Penambahbaikan Jarak Berpengaruh Pantas. Walaubagaimanapun, pengiraan kaedah ini tidak stabil, masih mengalami kesan penyorokan dan kesan limpahan, isu masa pengiraan dan tidak mengunakan titik-potong yang betul. Bagi mengatasi masalah ini, dicadangkan satu kaedah teguh versi baharu bagi jarak berpengaruh yang berasaskan penganggar $\sqrt{n}$ berpemberat konsisten pantas dan titik musnah tinggi. Keputusan menunjukkan bahawa kaedah yang dicadangkan sangat berjaya dalam mengenalpasti cerapan berpengaruh dengan masa pengiraan paling singkat, kesan limpahan paling kecil dan tiada kesan penyorokan berbanding dengan Kaedah Jarak Berpengaruh dan Kaedah Penambahbaikan Jarak Berpengaruh Pantas.

Penganggar M-teritlak (GM6) mempunyai beberapa kelemahan seperti ia nya menurunkan pemberat bagi titik tuasan tinggi baik dan mengambil masa pengiraan yang panjang. Bagi mengatasi masalah ini, diperkenalkan Penganggar M-teritlak yang berasaskan reja teritlak student tertambahbaik (GM-FIMGT). Walaubagaimanapun, kaedah ini menggunakan indeks set kesaksamaan dalam pengiraan jarak Mahalanobis teguh, yang diketahui tidak stabil dan masih mempunyai kesan penyorokan dan kesan limpahan. Oleh itu, satu penganggar baharu teguh GM-RFIID yang menggabungkan fungsi pemberat baharu yang dibina dari ukuran jarak berpengaruh versi baharu telah dibangunkan. Hasil kajian menunjukkan bahawa GM-RFIID adalah lebih cekap daripada penganggar GM-FIMGT dan penganggar GM6.

Penganggar Jackknife Ridge Teguh MM dan Jackknife Ridge GM2 dibangunkan untuk menyelesaikan masalah multikolinearan dengan kehadiran titik terpencil. Walaubagaimanapun, kedua-dua penganggar ini masih kurang cekap kerana mereka mempunyai masa pengiraan yang panjang dan tidak berjaya sepenuhnya untuk menangani titik tuasan tinggi. Seterusnya, regresi teguh Jackknife Ridge berasaskan GM-RFIID telah dibangunkan. Keputusan menunjukkan bahawa penganggar yang dicadangkan ini mengatasi kaedah sedia ada yang dibincangkan dalam tesis ini.

Tidak banyak penyelidikan telah dijalankan untuk menangani isu multikolinearan yang disebabkan oleh titik tuasan tinggi. Penghapusan suspek titik tuasan tinggi daripada analisis dengan menggunakan Kaedah Potensi Teritlak, Kaedah Kuasadua Terkecil Terpangkas- Potensi Teritlak Teguh Berdiagnostik-M dan Kaedah MM- Potensi Teritlak Teguh Berdiagnostik, adalah tiga kaedah yang terdapat dalam kesusteraan. Walaubagaimana pun, ketiga-tiga kaedah itu kurang cekap. Oleh itu, kami mencadangkan kaedah GM-RFIID bagi mengatasi masalah multikolinearan yang disebabkan titik tuasan tinggi. Keputusan menunjukkan bahawa kaedah GM-RFIID adalah lebih unggul berbanding dengan kaedah sedia ada yang dibincangkan dalam tesis ini.

ACKNOWLEDGEMENTS

My deepest thanks go out to Prof. Dr Habshah Midi, my supervisor, for all of the lessons and advice and inspiration she has provided for me during my duration of the study. Her unwavering encouragement and trust throughout my studies have been really helpful to me. Without her help and encouragement, I would not have been able to finish this dissertation, and I feel very lucky to had her as a teacher. I am also very thankful for the help, and advice, I got from Associate Prof. Dr. Jayanthi A/P Arasan, Dr. Mohd Shafie Bin Mustafa, and Dr. Hassan S. Uraibi, who were all on my committee.

I'd want to express my gratitude to every one of the Institute for Mathematical Research's employees for their assistance and support in one way or the other.

I would like to extend my heartfelt thanks to my parents, father TALIB HENDI and my mother for their support, encouragement, and prayers for me over the years, to also my brothers, sisters without their support this study wouldn't have been possible, and my thanks for relative for their moral support and prayers. Additionally, I appreciate the support and cooperation I have received from the Institute for Mathematical Research (INSPEM) and UPM as a whole.

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

Habshah binti Midi, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Jayanthi a/p Arasan, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Mohd Shafie bin Mustafa, PhD
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Member)

Hassan S. Uraibi, PhD
Senior Lecturer
Faculty of Business and Management
AL-Qadisiyah University, Iraq
(Member)

ZALILAH MOHD SHARIFF, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 14 March 2024

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

BIF	Bounded Influence Function
BLPs	Bad Leverage Points
BLUE	Best Linear Unbiased Estimators
BP	Breakdown Points
CD	Cooks Distance
CIO	collinearity influential observations
CVIF	Classical Variance Inflation Factor
DFFITS	Cook's distance and difference in fitted values
DGK	The Devlin, Gnanadesikan and Kettenring estimator
DRGP	Diagnostics Robust Generalized Potential
DRGP(ISE)	Diagnostic Robust Generalized Potential based on Index Set Equality
DRGP(MVE)	Diagnostic Robust Generalized Potential based on Minimum Volume Ellipsoid
DRGP(RFCH)	Diagnostic Robust Generalized Potential based on Reweighted Fast Consistent and High Breakdown Estimators
FCH	Fast Consistent and High breakdown
FIID	Fast Improvised Influential Distance
FIMGT	Fast Improvised Generalized Studentized Residual
FMGt	Fast Modified Generalized Studentized Residual
GDFF	Generalized Distance and Difference in Fitted values
GLO	Good Leverage Observation
GLPs	Good Leverage Points
GLV	Generalized Leverage Value
GM	Generalized M-estimator
GM-DRGP	Generalized M-estimator based on Diagnostic Robust Generalized Potential

GM-FIMGT	Generalized M-estimator based on Fast Improvised Generalized Potential
GM-RFIID	Generalized M-estimator based on RFIID
GP	Generalized Potential
GSR	Generalized Studentized Residual
GUM	Group Union Method
HBDP	high breakdown point
HLCEO	High leverage collinearity-enhancing observations
HLCRO	High leverage collinearity- reducing observations
HLPs	High Leverage Points
ID	Influential Distance
IF	Influence function
IO	Influential Observations
ISE	Index Set Equality
JRR	Jackknife Ridge Regression
LM	Lagrange Multiplier
LMS	Least Median Squares
LS	Least Squares
LTS	Least Trimmed Squares
MAD	Median Absolute Deviation
MB	Median Ball
MCD	Minimum Covariance Determinant
MD	Mahalanobis Distance
MM	Modify of M-estimator
MVE	Minimum Volume Ellipsoid
NMAD	Normal Median Absolute Deviation
OLS	Ordinary Lease Squire

PCE	Percentage of Change Estimator
RFCH	Reweighted Fast Consistent and High Breakdown Estimators
RFIID	Robust and Fast Improvised Influential Distances
RJGM-RFIID	Robust Jackknife Based on GM-RFIID
RJGM2	Robust Jackknife Based on GM2
RJMM	Robust Jackknife Based on MM-Estimator
RMD	Robust Mahalanobis Distance
RMD(ISE)	RMD based on Index Set Equality
RMD(MVE)	Robust Mahalanobis Distance based on Minimum Volume Ellipsoid
RO	Regular Observations
RR	Ridge Regression
RVIF(MM)	Robust Variance Inflation Factor based on MM-Estimator
TS	Three-Sigma
VIF	Variance Inflation Factor
VO	Vertical Outlier
WRR	Weighted Ridge Regression
WT	White test

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Regression analysis is a vital statistical approach for examining the linear relation between a response variable and one or more predictor variables. Sir Francis Galton first introduced it in the nineteenth century. There are multiple techniques available for modelling and analyzing variables in linear regression. One widely adopted technique is the ordinary least squares (OLS), which was introduced by Legendre and Gauss (Maronna et al., 2006) due to its simplicity and computational efficiency. The OLS is easy to use as it is available in most of the statistical software like SPSS, SAS and MINITAB. The OLS estimates are obtained by minimizing the sum of squared errors. It has many attractive features under normality assumption of regression errors, not only in parameters estimation but also in testing of hypothesis. The Gauss-Markov theorem states that under certain conditions, the OLS method provides the Best Linear Unbiased Estimator (BLUE) for the parameters of a linear model, specifically, the estimator has the smallest variance among those that are unbiased. Thanks to its advantageous properties, such as a closed-form solution and ease of computation, OLS is extensively applied in various fields of study, including applied sciences and engineering.

Nonetheless, many not aware that the OLS estimates suffer a huge setback when outliers are present in the data. According to Belsley et al. (1980), outliers refer to observations that have the greatest effect on the calculated values of various estimates, either alone or in combination with multiple other points. Additionally, Hawkins (1980) described an outlier as one that differs so significantly from the others that it raises concerns that it was caused by a variety of factors. “An outlier is an observation that, because it is unusual and/or unjustified, deviates decisively from the overall behavior of experimental data in regards to the criterion studied,” as explained by Muñoz-Garcia et al. (1990). As per Barnett and Lewis (1994), outliers are points that are significantly different from the bulk of observations included in a data collection.

In regression analysis, outliers can be classified into three main categories (residual outliers, vertical outliers, and high leverage points). Residual outliers refer to observations that have a large residual while vertical outliers refer to observations that are extreme or are outlying in the Y-space. High leverage points (HLPs) are those observations that fall far from the majority of the explanatory variables or are outlying in the X-space (see Figure 1.1 and Figure 1.2).

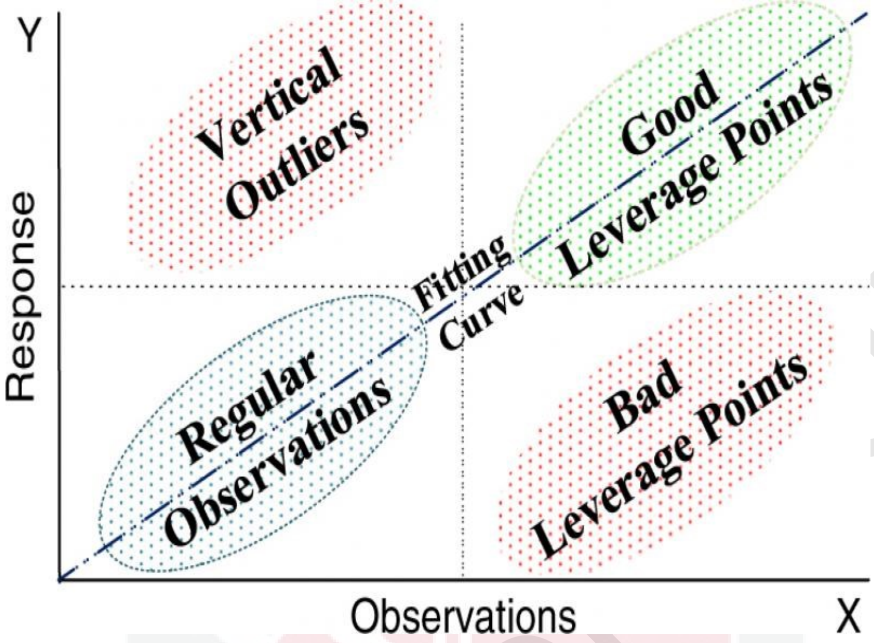

Figure 1.1 : Outliers in Y-Direction and X-Directions

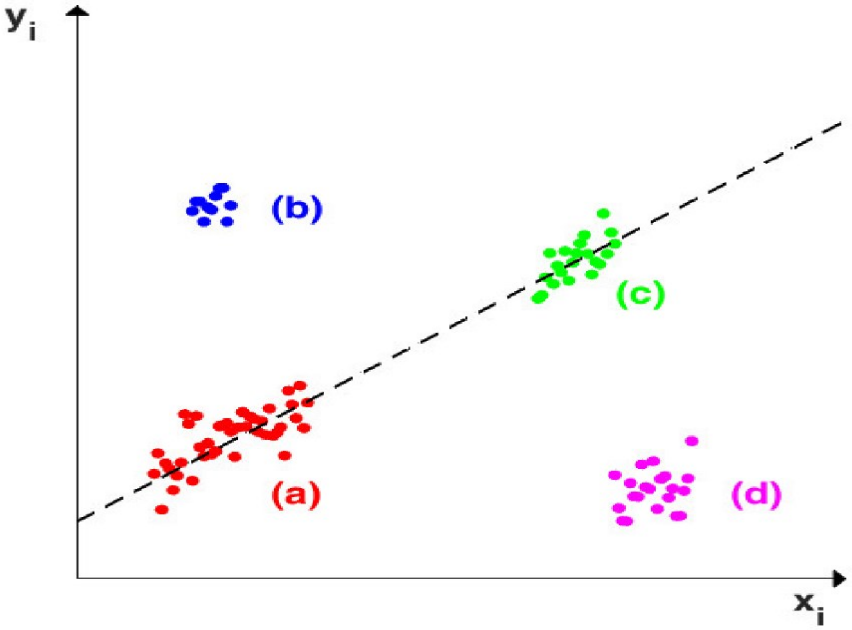

Figure 1.2 : Example of Simple Linear Regression Showing (a) Regular Observations, (b) Vertical Outliers, (c) Good Leverage Points, (d) Bad Leverage Points

According to Hampel et al. (1986), a routine data set typically contains about 1-10 % outliers, and even the highest quality data set cannot be guaranteed to be free of outliers. One immediate consequence of the presence of outliers may cause apparent non-normality (Huber, 1973). Since most of the statistical analysis are based on normality assumption, the violation of this assumption may lead to invalid inferential statements and inaccurate predictions. Evidences are now available in the literatures that the presence of outliers have an adverse effect on the computed values of various estimates (Rousseeuw, 1985; Imon & Khan, 2003; Midi et al., 2009; Riazoshams et al., 2010; Bagheri et al., 2012, Zahariah et al., 2021, Rashid et al., 2022). Rashid et al. (2021) pointed out that among the three types of outliers, leverage point has the most detrimental effect on the computed values of various estimates which leads to misleading conclusion about the fitted regression model.

It is now evident that in the presence of multicollinearity, the OLS can result in a very poor estimate. Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. According to Bagheri et al. (2012) and Bagheri and Midi (2015), multicollinearity can cause the OLS estimates to have large variances which lead to inaccurate prediction. The problem is further complicated when both outliers and multicollinearity are present in a data. Robust methods alone cannot remedy the simultaneous problems of multicollinearity and outliers.

Hence robust statistical methods that are able to reduce/eliminate the effect of multicollinearity and outliers should be used as an alternative to the classical method. Thus, in this thesis, we will develop several alternatives robust methods to deal with multicollinearity and outliers. Several basic concepts and some commonly used methods need to be introduced in the following sections.

1.2 Mahalanobis Distance (MD)

Mahalanobis Distance (MD) measures how far each point is from the centroid of all points for the independent variables. It is initially introduced by Mahalanobis (1936), finds extensive application in multivariate analysis as a prominent metric for measuring the dissimilarity between two individual points within a dataset containing multiple variables (Filzmoser et.al, 2009).

Let present the i^{th} vector of independent variables as:

$$X_i' = (1, X_1, X_2, \dots, X_P) = (1, t_i),$$

where t_i is a $p-$dimensional row vector. The vector of the mean and the variance covariance matrix can be calculated, respectively as:

$\bar{t} = {}^{1}/_{n} \sum_{i=1}^{n} t_i$

$C = \left(\frac{1}{n-1}\right) \sum_{i=1}^{n} (t_i - \bar{t}\,)\ (t_i - \bar{t}\,)',$

The (MD) for each observation is defined as follows:

$$MD_i = \sqrt{\left(t_i - T(X)\right)' C(X)^{-1} \left(t_i - T(X)\right)} \qquad i = 1,2,\dots,n, \qquad (1.1)$$

and Mahalanobis (squared) distance is given by,

$$MD_i^2 = \left(t_i - T(X)\right)' C(X)^{-1} \left(t_i - T(X)\right), \quad i = 1, 2, 3, \dots, n$$

where $T(X)$ is the mean vector $(\,\bar{t}\,)$ and $C(X)$ is the variance covariance matrix (C).

1.3 Minimum Volume Ellipsoid (MVE)

The Minimum Volume Ellipsoid (MVE) represents another robust multivariate estimator. Rousseeuw (1985) stated that the center of the minimal volume ellipsoid covers at least half of the h points of X, where h can be taken as equal to $\left[\frac{n}{2}\right] + 1$. This ellipsoid can serve as an effective covariance predictor. However, in practical applications, considering all possible subsets of data becomes infeasible. Midi et al. (2009) proposed an approach to compute the MVE by first drawing a subsample of (p + 1) distinct observations, indexed by $J = (i_1, i_2, \dots i_p)$. The mean and covariance matrix are then determined as follows:

$$\bar{x}_j = \frac{1}{p+1} \sum_{i \in J} x_i \text{ and } C_j = \frac{1}{p} \sum_{i \in J} (x_i - \bar{x})^T (x_i - \bar{x}).$$

Assuming that the C_j is non-singular, the corresponding ellipsoid needs to be adjusted to precisely enclose h points, which involves the computation of $m_J^2 = Med(x_i - \bar{x}_J) C_J^{-1} (x_i - \bar{x}_J)^T$. The resulting ellipsoid's volume, denoted by $m_J^2 C_J$, is proportional to $\sqrt{det(\, m_J^2 C_J)} = \sqrt{det(\, C_J)} (m_J)^P$. This process is repeated for multiple J iterations to minimize the above determinant, and the corresponding values are obtained as follows:

$T(X) = \bar{x}_J$ and $C(X) = (x_{p,0.5}^2)^{-1}\, m_J^2 C_J$,where $x_{p,0.5}^2$ is the median of the Chi-square distribution with p degrees of freedom. This correction factor is required to attain the consistency for multivariate normal data.

1.4 Basic Notions

Robust estimators are designed to offer valuable information even when some of the parametric assumptions are violated. In the context of linear regression analysis, robust regression methods are employed to yield resistant estimates, ensuring stable results even in the presence of unusual observations in the dataset (for further details, refer to Huber, 1964; Hampel, 1974; Andrews, 1974; Ramsay, 1977; Simpson, 1995; Rousseeuw and Leroy, 1987; Wilcox, 2005; Maronna et al., 2006).

The primary objective of a robust estimator is to provide estimates based on the information contained in the majority of the data set. It seeks to fit a model that relies on the information from the most significant portion of the data. The fundamental properties used to assess the performance of robust estimators are efficiency, breakdown point, and bounded influence. These principles are briefly stated as follows:

1.4.1 Efficiency

Efficiency serves as a metric to gauge how effectively a robust method performs compared to the least squares method under basic assumptions. It can be expressed as a percentage, representing the ratio between the variance of the least squares fits on the clean data (without outliers) and the variance of the robust fit (Maronna et al., 2006). An efficient estimator is also known as the minimum variance unbiased estimator (MVUE), as it achieves the minimum variance among all parameter estimates. Precision is another vital aspect of an estimator, and it is quantified by its statistical efficiency. The statistical efficiency of an estimator depends on the assumed distribution. For example, the sample mean exhibits perfect efficiency of 100% when the distribution is normal, but its efficiency may vary with other distributions.

1.4.2 Breakdown Point

A robust approach aims to possess a high breakdown point, a crucial characteristic. The breakdown point (BP) represents the minimum percentage of contamination that can completely disrupt or collapse an estimator or estimating process (Hampel, 1974; Coakley and Hettmansperger, 1993). Conversely, even a small number of bad data points (outliers) can significantly distort an estimator. When an estimator has a high breakdown point, it can withstand a substantial number of outliers without the analysis collapsing. In practical terms, this means the estimate remains stable as long as less than 50% of the data are replaced with outlying observations, and the maximum attainable BP is thus 0.50 (Rousseeuw and Croux, 1993). Furthermore, to formally define breakdown in a finite sample, we can consider a sample of n data points as follows:

$$G = \{(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)\} .$$

If we assume T to be a regression estimator, we get the following vector of regression coefficients when we apply T to such a sample G:

$$T(G) = \hat{\beta}.$$

To obtain all possible corrupted samples G^T, any m of the original data points is replaced with arbitrary values, or also known as outliers. Therefore, the estimator $T's$ breakdown point at sample G is defined as

$$BP(T,G) = \min\left\{\frac{m}{n};\, \underset{G^T}{\text{SUP}}\|T(G^T) - T(G)\| \text{ is infinite}\right\}$$

in which the supremum is over all possible data matrix G, which contains $n-m$ observation and m contaminated points (Rousseeuw and Leroy, 1987; Maronna et al. 2006).

1.4.3 Bounded Influence Function

The bounded influence function (BIF), as demonstrated by Simpson (1995), exhibits robustness against high leverage points in the X space. Its primary purpose is to safeguard model estimators from the influence of outlier points within the X space. Additionally, the influence function (IF) is utilized to assess the robustness of an estimator concerning minor contamination levels, often serving as a means to determine whether the estimator possesses BIF.

The following expression represents the IF of an estimator T at a distribution F, calculated for points $x_0$within the sample space, provided the limit exists:

$$\text{IF}(x_0;T,F) = \lim_{\delta\to\infty} \frac{T((1-\delta)F + \delta\varphi_{x_0}) - T(F)}{\delta},$$

where φ_{x_0} denotes the probability distribution that puts all its mass in the point x_0 and δ represents the contamination amount. Moreover, it is vital to note that the influence function reflects the bias introduced by a few outliers at the point x_0 (Rousseeuw and Leroy, 1987; Simpson, 1995; Wilcox, 2005; Maronna et.al, 2006).

1.5 Problem Statement

As will be presented in Chapter 2 (Literature Review) many methods of identification of HLPs have been proposed. Among all methods that will be discussed, most of them are not very successful in detecting multiple HLPs. As a solution to this, Midi et al. (2009) established Diagnostic Robust Generalized Potential (DRGP) which is based on Minimum Volume Ellipsoid (MVE) in order to improve the detection rate of high leverage points. The DRGP(MVE) is very successful in the detection of HLPs. However, it was reported that the algorithm of DRGP(MVE) has complex computational form and has a long computational running time. The iterative computational form of

MVE makes DRGP(MVE) to consume a lot of time. Lim and Midi (2016) proposed another version of DRGP based on index set equality (ISE) denoted by DRGP(ISE) in order to reduce the computational complexity of DRGP(MVE) of Midi et al. (2009). Nonetheless, DRGP(ISE) is computationally not stable and still suffers from masking and swamping effects. This shortcoming has motivated us to improvise the existing DRGP by integrating the location and scatter estimators obtained from the Reweighted Fast Consistent and High breakdown (RFCH) estimators (Olive & Hawkins, 2010). The developed method is denoted as DRGP(RFCH). We also suggest using Q_n (estimate of scale) as robust alternative to scale estimate, instead of using NMAD (Midi et al.2009) in the computation of cut-off point for the proposed DRGP(RFCH) which is denoted as p_{ii} (p_{ii} refer to potential values), to improve the accuracy of the identification of HLPs. As reported by Rousseuw and Croux (1993), the efficiency of NMAD at normal data is only 37% as compared to median which has 64% efficiency. Furthermore, both Q_n and NMAD have desirable robustness properties (50% breakdown point and bounded influence). However, Q_n has better efficiency (82%) and does not depend on symmetry.

Influential observations will be discussed at length in Chapter 2 (Literature Review). Belsley et al. (2004) noted that influential observations (IOs) are those observations which either alone or together with several other observations have detrimental effect on the computed values of various estimates. Hence, it is very crucial to detect them before making any statistical analysis or inferential statements. Many methods have been developed to detect IOs that includes the “influential distance” (ID) of Nurunnabi et al. (2016) and fast influential distance (FIID) introduced by Midi et al. (2021b). The ID technique comprises three main stages. Initially, the first stage involves identifying suspected unusual observations through the Group Union Method (GUM). In the second stage, it detects high leverage points (HLPs) and vertical outliers (VOs). And in the third stage, it calculates the influential distance (ID). Overall, this method is highly effective in identifying influential observations (IOs). However, this method has a drawback concerning the first stage and when deciding to consider IOs. In the first stage, the technique employs the union of five different detection methods (standardized studentized residual, standardized LMS residuals, leverage values or hat matrix, Cook’s distance, and difference in fits) to identify the suspected unusual observations. Some of these detection methods have been reported to suffer from a high rate of masking and swamping, as discussed in Midi et al. (2009). Additionally, the computation time required for all these diagnostic methods is substantial. Hadi (1992) highlighted the critical importance of choosing the initial suspected unusual observations accurately, as it significantly affects the correct detection of the final influential observations (IOs). The suspected IOs of Nurunnabi et al. (2016) is then confirmed by sketching a confidence bound on the ID’s plot of GSR versus GLV and declared an observation is IO if it falls beyond its confidence bound. This approach will ultimately tend to declare more good observations as IOs and taking longer computational running time. To rectify this problem, Midi et al. (2021b) introduced an improved version of ID (FIID) that exhibits reduced computational running times. This version has demonstrated considerable success in detecting IOs with diminished swamping and masking effects. Despite these advancements, the efficiency of FIID in terms of computation running times and IO detection is still limited due to its reliance on index set equality (ISE) as a fundamental basis. Recent findings have indicated that ISE can be inherently unstable because the algorithm’s outcomes are contingent upon the initial subset, ‘h,’ that is selected. Midi et al. (2021b) illustrated that the final estimator of location and scatter matrix based on ISE is analogous to the minimum covariance determinant (MCD) when the same initial subset is employed. However, if a different initial subset is utilized, the results may significantly diverge. Another shortcoming of this method is the use of F distribution as cut-off point for FIID which is not suitable since the distribution of FIID is intractable. The weaknesses of this estimator have inspired us to develop an alternative method of identifying IOs named robust and fast influential distance denoted as RFIID. The DRGP (RFCH) is integrated in the algorithm of RFIID.

Many robust estimators that are resistance to outliers such as the M, MM, S, Least Trimmed Squares (LTS), Least Median of Squares (LMS) are developed to address the issue of outliers affecting parameter estimation methods. While some of these methods exhibit high efficiency and possess high breakdown points (HBDP), they lack bounded influence properties, in the sense that they cannot cope with HLPs. The GM6 estimator which is based on robust Mahalanobis distance (RMD) which uses minimum volume ellipsoid (MVE) as an initial of d–weight function (Rousseeuw, 1985) is noted to have bounded influence properties. However, as will be discussed in Chapter 2, the GM6 has several drawbacks due to using MVE. The computation running time for MVE is very intensive. The MVE is noted to swamp some low leverage as high leverage points. Besides, the RMD which is based on MVE attempts to identify high leverage points without taking into consideration whether they are good or bad leverage points. Hence, the GM6-MVE considers the good leverage points as bad leverage points and its efficiency tends to decrease as the number of good leverage points increases. The weaknesses of GM6-MVE have prompted Midi et al. (2021a) to put forward another version of GM estimator which is called the Fast GM estimator which is based on Improvised Generalized MT (FIMGT). It is denoted as GM-FIMGT estimator. The Fast GM estimator utilizing high breakdown point S-estimator as an initial estimate and using d −weight function based on Fast Improvised Generalized MT (FIMGT). It has been shown that the GM-FIMGT is more efficient than the GM6 estimator. The only shortcoming of this method is that the FIMGT is based on index set equality (ISE) whereby it is computationally not stable. Hence, their work has motivated us to propose a new version of GM based on robust and fast influential distance RFIID that is denoted as GM-RFIID. The RFIID is incorporated in the algorithm of GM-RFIID.

This thesis also addresses the issue of multicollinearity in multiple linear regression models in the presence of HLPs. The problem of multicollinearity can stem from various sources, such as the data collection method employed, model constraints, model specification, and an overly defined model (Montgomery et al., 2021). It is important to note that the OLS estimates are much affected by the existence of multicollinearity in which it occurs when two or more independent variables are highly correlated. Multicollinearity results in undesirable consequences and relying on the OLS estimator can result in a wrong sign problem of regression coefficients, lead to erroneous interpretation of regression coefficient, causes regression estimates to have unduly large variances and inaccurate prediction. Many remedial techniques to address the adverse effects of multicollinearity will be discussed in Chapter 2 (Literature Review). The ridge regression estimator (RR) is the commonly used method to remedy the problem of multicollinearity (Hoerl ,1962; Hoerl and Kennard ,1970b; Hoerl et al.,1975; Marquard and Snee,1975). However, Batah et al. (2008) and Akdeniz and Akdeniz, (2012) noted that the RR estimator is significantly biased. Therefore, Singh et al. (1986) proposed an unbiased ridge estimator based on the Jackknife method. Unfortunately, RR techniques alone and robust method alone are not adequate enough to address the complicated problems of multicollinearity and outliers (Midi and Zahari, 2007; Alguraibawi et al., 2015; Zahariah et al., 2021). As a solution to this shortcoming, Jadhav and Kashid (2011) suggested the use of Jackknife ridge M-estimator to overcome multicollinearity and outliers in the Y direction. Alguraibawi et al. (2015) proposed a combination of the Jackknife Ridge Regression (JRR) with MM-estimator and called Robust Jackknife MM estimator, denoted as RJMM. They also proposed combining the Jackknife Ridge Regression with the GM2-estimator of Bagheri (2011) and denoted the estimator as RJGM2. As will be discussed in Chapter 2 (Literature Review), the GM2 is noted to have several drawbacks due to using MVE in the computation of RMD. It takes longer computational running times and its efficiency decreases as the number of good leverage points increases. Thus, this drawback has motivated us to establish Robust Jackknife Ridge Regression based on GM-RFIID denoted as RJGM-RFIID.

Many are not aware that beside data collection strategy, model constraints, model specification, and an over-defined model, the existence of HLPs is another prime source of multicollinearity (Imon and Khan, 2003). Bagheri et al. (2012) noted that multicollinearity caused by HLPs is related to collinearity-influential observation, in which it can alter multicollinearity pattern of a data. Specifically, the high leverage collinearity-enhancing observations (HLCEO) that induce multicollinearity in a data set is the prime source of multicollinearity. Regardless of the source, detecting multicollinearity is very important, so that correct remedial measure is taken up to obtain efficient parameter estimates. Corrective techniques such as ridge regression and jackknife ridge regression are ineffective when multicollinearity is caused by the presence of HLPs. Very scarce work is devoted to this issue. Since multicollinearity is caused by HLPs, the only solution to this problem is by using robust method that able to reduce the effect of HLPs. Imon and Khan (2003) suggested deletion of suspect HLPs from the analysis by using generalized potentials (GP). Nonetheless, Midi et al. (2009) noted that the GP approach is not very successful in identifying HLPs. Bagheri and Midi (2009) proposed using their developed GM-DRGP-LTS and DRGP-MM estimators to remedy multicollinearity caused by the presence of HLPs. However, those estimators are still not very efficient because their derivation involve RMD based on MVE which have been mentioned previously to have several drawbacks. Hence, their work has inspired us to propose the new version of GM estimator (GM-RFIID) developed in Chapter 5, to remedy the problem of multicollinearity caused by HLPs.

1.6 Research Objectives

The objective of our research can be outlined as follows:

1. To develop a fast-diagnostic technique for the detection of high leverage points in multiple linear regression model (DRGP (RFCH)).
2. To develop Fast Improvised Method of Identification of Influential Observations in multiple linear regression model (RFIID).
3. To develop a fast version of high breakdown, high efficiency, bounded influence GM estimator for multiple linear regression model (GM-RFIID).

4. To develop Robust Jackknife Ridge Regression based on GM-RFIID to simultaneously remedy the multicollinearity problems and outliers.
5. To apply GM-RFIID for handling multicollinearity problems caused by HLPs.

1.7 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. Multiple linear regressions are predominantly fitted using the OLS method because of tradition and ease of computation if the underlying assumptions are hold. Unfortunately, the OLS estimate is not robust against outliers and influential observations. Its performance worsens when both multicollinearity and outliers are present in a data set. Moreover, there is evidence that HLPs is a new source for multicollinearity. Hence the scope of this thesis concentrates on the establishment of diagnostic and robust methods in multiple linear regression model.

The most challenging aspect in this thesis evolves around creating the necessary programming codes. Finding dataset that suited the particular problems at hand is also another challenging task. In most of the chapters, we only report the simulation study for few numbers of parameters (p =2,3,4,5). The results of another p are consistent.

1.8 Outline of the Thesis

This thesis consists of eight chapters based on the research objectives and scope of the study. Chapters of the thesis are structured in a way that makes it clear what the objectives are and how they relate to each other.

Chapter Two discusses the literature reviews on the diagnostic methods of high leverage points and vertical outliers in the multiple linear regression. This chapter also reviews some methods on the identification of influential observations. Some existing robust regression methods for parameter estimation in the presence of HLPs and vertical outliers are also presented. Finally, the literature reviews on multicollinearity with its consequences and multicollinearity sources with its usual detection and estimation techniques.

Chapter Three is mainly dedicated to develop a Fast Diagnostic Technique for the Detection of High Leverage Points in multiple linear regression model, named as DRGP(RFCH). The performance of DRGP(RFCH) is evaluated using real data sets and simulation studies.

Chapter Four discusses the new method for the identification of multiple influential observations termed Robust and Fast Improvised Influential Distance (RFIID). Several well-referred real data set and Monte Carlo Simulation study are conducted to evaluate the performance of the proposed method.

Chapter Five presents the development of the GM-estimator in the linear regression, denoted by GM-RFIID. Real data set and Monte Carlo simulations are presented to assess the performance of the proposed method.

Chapter Six discusses the establishment of the new robust Jackknife Ridge Regression based on GM-RFIID to simultaneously remedy the multicollinearity problems and outliers. Real data set and Monte Carlo simulations are presented to evaluate the performance of the proposed method.

Chapter Seven discusses the issue of multicollinearity arising from the presence of highly leveraged points (HLPs) in a data set. This type of phenomena is tackled by utilizing the GM-estimator. The discussion is focused on using GM-RFIID using GM-

To evaluate the effectiveness of the proposed technique, a Monte Carlo simulation study and numerical examples are conducted.

Chapter Eight presents a summary, conclusions, and recommendations for future research.

REFERENCES

Akdeniz Duran, E., & Akdeniz, F. (2012). Efficiency of the modified jackknifed Liu-type estimator. *Statistical Papers*, 53, 265-280.

Alguraibawi, M., Midi, H., & Imon, A. H. M. (2015). A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering,* 2015.

Alguraibawi, M., Midi, H., & Rana, S. (2015). Robust Jackknife Ridge Regression to Combat Multicollinearity and High Leverage Points in Multiple Linear Regressions. *Econ. Comput. Econ. Cybern. Stud. Res*, 4, 305-322.

Alkenani, A., & Yu, K. (2013). A comparative study for robust canonical correlation methods. *Journal of Statistical Computation and Simulation,* 83(4), 692-720.

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.

Askin, R. G. & Montgomery, D. C. (1980). Augmented robust estimators, *Technometrics*. 22(3), 333-341.

Askin, R. G., & Montgomery, D. C. (1984). An analysis of constrained robust regression estimators. *Naval research logistics quarterly*, 31(2), 283-296.

Bagheri, A. (2011). *Robust Estimation Methods and Robust Multicollinearity Diagnostics for Multiple Regression Model in the Presence of High Leverage Collinearity-Influential Observations* (Doctoral dissertation submitted to the School of Graduate Studies, University Putra Malaysia).

Bagheri, A., & Midi, H. (2009). Robust estimations as a remedy for multicollinearity caused by multiple high leverage points. *Journal of Mathematics and Statistics*, 5(4), 311.

Bagheri, A., & Midi, H. (2012). On the performance of the measure for diagnosing multiple high leverage collinearity-reducing observations. *Mathematical Problems in Engineering*, 2012.

Bagheri, A., & Midi, H. (2015). Diagnostic plot for the identification of high leverage collinearity-influential observations. *SORT-Statistics and Operations Research Transactions,* 39(1), 51-70.

Bagheri, A., Habshah, M. (2011). On the performance of robust variance inflation factors. International. *Journal of Agricultural and Statistics Sciences,* 7, 31-45.

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.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data*. Wiley. New York.

Batah, F. S. M., Ramanathan, T. V., & Gore, S. D. (2008). The efficiency of modified jackknife and ridge type regression estimators: a comparison. *Surveys in Mathematics and its Applications*, 3, 111-122.

Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2), 147-185.

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

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

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350), 428-434.

Brown, P. J. (1977). Centering and scaling in ridge regression. *Technometrics*, 19(1), 35-36.

Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical science*,1(3), 379-393.

Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*, Wiley, New York.

Chatterjee, S., & Hadi, A. S. (2006). *Regression Analysis by Example*, New Jersey: John Willey & Sons.

Coakley, C. W., & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association,* 88(423), 872-880.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374), 354-362.

Eledum, H., & Zahri, M. (2013). Relaxation method for two stages ridge regression estimator. *International Journal of Pure and Applied Mathematics,* 85(4), 653-667.

Filzmoser, P., Liebmann, B., & Varmuza, K. (2009). Repeated double cross validation. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 23(4), 160-171.

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

Groβ J. (2003). *Linear Regression- Lecturer Notes in Statistics*, Springer Verlag Berlin Heidelberg.

Groβ, J. (2003). Linear regression. *Lecture Notes in Statistics,* 175, 12.

Gujarati, D. N. (2003). *Basic Econometrics*. fourth edition McGraw-Hill. *New York.*

Gunst, R. F., & Mason, R. L. (1980). *Regression Analysis and Its Application: A Data Oriented Approach*. New York: Marcel Dekker.

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

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

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association,* 69(346), 383-393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., & Stahel, W. A. (1986). Robust statistics: the approach based on influence functions. Wiley-Interscience; New York.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. (2011). Robust Statistics: The Approach based on Influence Functions. Hoboken, Ney Jersey: John Wiley & Sons, Inc.

Hawkins, D. M, Bradu, D., & Kass, G. V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics,* 26 (3), 197–208.

Hawkins, D. M. (1980). *Identification of outliers*. Chapman and Hall, London.

Hekimoğlu, S., & R. C. Erenoglu (2013). "A new GM-estimate with high breakdown point." *Acta Geodaetica et Geophysica,*4(48): 419-437.

Hill, R. W. (1977). *Robust regression when there are outliers in the carriers*. Unpublished Ph.D. thesis. Harvard University, Boston, MA.

Hill, R. W. & 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., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician,* 32(1), 17-22.

Hocking, R.R., & Pendelton, O.J. (1983). The regression dilemma, *Communications in Statistics-Theory and Methods* 12*(5)*: 497-527.

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: applications to non-orthogonal problems. Technometrics. 12(1):69-82.

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, 4(2), 105-123.

Hoerl, A.E. (1962), "Application of Ridge Analysis to Regression Problems," *Chemical Engineering Progress,* 58, 54-59.

Hossain, M. G., Zyroul, R., Pereira, B. P., & Kamarul, T. (2012). Multiple regression analysis of factors influencing dominant hand grip strength in an adult Malaysian population. *Journal of Hand Surgery (European Volume),* 37(1), 65-70.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics,* 35(1), 73-101.

Huber, P. J. (1973). Robust regression: asymptotic, conjectures and Monte Carlo. *The Annals of Statistics,* 1(5), 799-821.

Huber, P. J., & Ronchetti, E. M. (1981). *Robust statistics.* john Wiley & sons: New York, 1(1).

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

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

Huber, P. J. (2005). *Robust statistics*, John Wiley & Sons.

Imon, A. H. M. R. (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies*, 3, 207-218.

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

Imon, A. H. M.R. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied statistics*, 32(9), 929-946.

Ismaeel, S. S., Midi, H., & Sani, M. (2021). Robust Multicollinearity Diagnostic Measure for Fixed Effect Panel Data Model. *Malaysian J. Fundam. Appl. Sci.*, 17(5), 636-646.

Jadhav N. H. & 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, M. D., & Imon, A. H. M. R. (2002). High leverage point: Another source of multicollinearity. *Pakistan Journal of Statistics-All Series-*, 18(3), 435-448.

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

Krasker, W. S., & R. E. Welsch (1982). "Efficient bounded-influence regression estimation." *Journal of the American Statistical Association,*77(379): 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., & Arthur, J. L. (1990). *Robust Regression; Analysis and Applications*, INC: Marcel Dekker.

Leroy, A., & Rousseeuw, P. (1987, August). A Robust Scale Estimator Based on the Shortest Half. In *17th European Meeting of Statisticians. Thessaloniki* ,24-28.

Li, G., & 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.

Lim, H. A., & Midi, H. (2016). Diagnostic Robust Generalized Potential Based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. *Computational Statistics* 3(31),859-877.

Lopuhaa, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *Annals of Statistics*,27(5), 1638-1665.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *National Institute of Science of India,* 2(1), 49-55.

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

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

Marquardt, D. W., & Snee, R. D. (1975). Ridge Regression in Practice. *The American Statistician*, 29(1), 3–20. https://doi.org/10.2307/2683673.

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

McDonald, G. C., & Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 407-416.

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

Midi, H., & Ismaeel, S. S. (2018). Fast improvised diagnostic robust measure for the identification of high leverage points in multiple linear regression. *Journal of Statistics and Management Systems,* 21(6), 1003-1019.

Midi, H., & Shabbak, A. (2011). Robust multivariate control charts to detect small shifts in mean. *Mathematical Problems in Engineering*, 2011.

Midi, H., & Zahari, M. (2007). A simulation study on ridge regression estimators in the presence of outliers and multicollinearity. *Jurnal Teknologi*, 47(1).

Midi, H., Bagheri, A., & Imon, A. H. M. R. (2011). A Monte Carlo simulation study on high leverage collinearity-enhancing observation and its effect on multicollinearity pattern. *Sains Malaysiana*, 40(12), 1437-1447.

Midi, H., Ismaeel S. S., & Arasan J. (2018). On the performance of fast Robust Variance Inflation factor based on index set equality. *Journal of Engineering and Applied Sciences.* 13(16), 6634-6638.

Midi, H., Ismaeel, S. S., Arasan, J., & Mohammed, M. A. (2021a). Simple and Fast Generalized-M (GM) Estimator and Its Application to Real Data Set. *Sains Malaysiana,* 50(3), 859-867.

Midi, H., Norazan, M. R., & 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.

Midi, H., Rana, S., & Imon, A. H. M. (2014). Tow-step robust estimator in heteroscedastic regression model in the presence of outliers. *Economic Computation & Economic Cybernetics Studies & Research,* 48(3),255-272.

Midi, H., Sani, M., Ismaeel, S. S., & Arasan, J. (2021b). Fast Improvised Influential Distance for the Identification of Influential Observations in Multiple Linear Regression. *Sains Malaysiana,* 50(7), 2085-2094.

Montgomery, D. C., Peck, E. A., & Viving, G.G. (2021). *Introduction to linear regression Analysis*. 4rd edition. New York: John Wiley and sons.

Muñoz-Garcia, J., Moreno-Rebollo, J., & Pascual-Acosta, A. (1990). Outliers: A formal approach. *International Statistical Review/Revue Internationale De Statistique,* 58(3), 215-226.

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

Nurunnabi, A. A. M., Nasser, M., & Imon, A. H. M. R. (2016). Identification and classification of multiple outliers, high leverage points and influential observations in linear regression. *Journal of Applied Statistics*, 43(3), 509-525.

Olive, D. J., & Hawkins, D. M. (2008). High breakdown multivariate estimators. *Preprint, see (www. math. siu. edu/olive/preprints. htm)*.

Olive, D.J., & Hawkins. (2010). *Robust multivariate location and dispersion*. Retrieved,December,fromhttps://www.researchgate.net/profile/David_Olive2/publication/228434748_Robust_multivariate_location_and_dispersion/links/02bfe51015be5c88ca000000.pdf.

Peña, D., & Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society: Series B (Methodological),* 57(1), 145-156.

Pfaffenberger, R. C., & Dielman, T. E. (1985). A comparison of robust ridge estimators. In *Proceedings of the American Statistical Association Business and Economic Statistics Section*, 631-635.

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

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.

Rashid, A. M., Midi, H., Slwabi, W. D., & Arasan, J. (2021). An efficient estimation and classification methods for high dimensional data using robust iteratively reweighted SIMPLS algorithm based on nu-support vector regression. *IEEE Access*, 9, 45955-45967.

Rashid, A.M., Midi, H., Dhhan, W. & Arasan, J. (2022). Detection of outliers in high-dimensional data using nu-support vector regression. *Journal of Applied Statistics*. 49(10), 2550-2569.

Riazoshams, H., & Midi, H. (2014). Robust Nonlinear Regression: Case Study for Modeling the Greenhouse Gases, Methane and Carbon Dioxide Concentration in Atmosphere. *Malaysian Journal of Mathematical Sciences*, 8(S), 173-184.

Riazoshams, H., & Midi, H. B. (2016). The performance of a robust multistage estimator in nonlinear regression with heteroscedastic errors. *Communications in Statistics-Simulation and Computation,* 45(9), 3394-3415.

Riazoshams, H., Midi, H., & Sharipov, O. S. (2010). The performance of robust two-stage estimator in nonlinear regression with autocorrelated error. *Communications in Statistics-Simulation and Computation,*39(6), 1251-1268.

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.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association*, 79(388), 871-880.

Rousseeuw, P. J. (1985). Multivariate Estimation with High Breakdown Point. *Mathematical statistics and applications,* 8, 283-297.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424), 1273-1283.

Rousseeuw, P. J., & Driessen, K. V. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics,* 41(3), 212-223.

Rousseeuw, P., & Van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of American Statistical Associations*, 85(411): 633-639.

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle & D. Martin (Eds.), Robust and nonlinear time series analysis ,256-272. New York, NY: Springer.

Rousseeuw, P.J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley & Sons.

Ruppert, D., & Simpson, D. G. (1990). Unmasking multivariate outliers and leverage points: comment. *Journal of the American Statistical Association*, 85(411), 644-646.

Salleh, R. (2013). A robust estimation method of location and scale with application in monitoring process variability *(Doctoral dissertation, Universiti Teknologi Malaysia).*

Simpson, D. G., Ruppert, D., & Carroll, R. J. (1992). On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association,* 87(418), 439-450.

Simpson, J. R. (1995). New methods and comparative evaluations for robust and biased-robust regression estimation. Unpublished PhD thesis, Arizona State University, the United State of America.

Singh, B., Chaubey, Y. P., & Dwivedi, T. D. (1986). An almost unbiased ridge estimator. *Sankhyā: The Indian Journal of Statistics, Series B*, 342-346.

Stromberg, A. J., & Ruppert, D. (1992). Breakdown in nonlinear regression. *Journal of the American Statistical Association*, 87(420), 991-997.

Stromberg, A. J., Hössjer, O., & Hawkins, D. M. (2000). The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association,* 95(451), 853-864.

Uraibi, H. S., Midi, H., & Rana, S. (2017). Selective overview of forward selection in terms of robust correlations. *Communications in Statistics-Simulation and Computation,* 46(7), 5479-5503.

Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35(4), 234-242.

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

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 48,817-838.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing (Statistical Modeling and Decision Science)*. Academic press.

Yohai, V. J., & Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American statistical association,* 83(402), 406-413.

Yohai, V. J., & Zamar, R. H. (1988). High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American statistical association*, 83(402), 406-413.

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

Zahariah, S., Midi, H., & Mustafa, M. S. (2021). An improvised SIMPLS estimator based on MRCD-PCA weighting function and its application to real data. *Symmetry*, 13(11), 2211.

Zhang, J., Olive, D., & Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability,* 1(2), 119.