ESTIMATION OF PARAMETERS AND OUTLIER DETECTION IN REPLICATED LINEAR FUNCTIONAL RELATIONSHIP MODEL

AZURAINI BINTI MOHD ARIF

INSTITUTE FOR ADVANCED STUDIES
UNIVERSITI MALAYA
KUALA LUMPUR

2023

ESTIMATION OF PARAMETERS AND OUTLIER DETECTION IN REPLICATED LINEAR FUNCTIONAL RELATIONSHIP MODEL

AZURAINI BINTI MOHD ARIF

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE FOR ADVANCED STUDIES
UNIVERSITI MALAYA
KUALA LUMPUR

2023

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **Azuraini binti Mohd Arif**

Registration/Matric No: **17027502/1 (HHC140013)**

Name of Degree: **Doctor of Philosophy (Ph.D.)**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Estimation of Parameters and Outlier Detection in Replicated Linear Functional Relationship Model

Field of Study: **Statistics**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ESTIMATION OF PARAMETERS AND OUTLIER DETECTION IN REPLICATED LINEAR FUNCTIONAL RELATIONSHIP MODEL

ABSTRACT

The thesis focuses on parameter estimation especially in the presence of outliers, outlier detection and grouping procedures in a linear functional relationship model (LFRM). There are two categories of LFRM: the unreplicated and replicated model. The study starts by modifying the maximum likelihood estimation method in unreplicated LFRM when the ratio of error variances is equal to one. A robust slope estimator namely the modified maximum likelihood estimation method is proposed. Results from simulation studies show that the modified maximum likelihood estimation method is outlier resistant and performs well than the traditional maximum likelihood estimation method. Then, an improvement on the estimation of the parameters by introducing balanced replicated observations in the LFRM when there is no information about the ratio of error variances is proposed. The estimation of parameters using maximum likelihood estimation method along with the variance-covariance matrix using the Fisher Information matrix is derived. Based on the simulation studies, the estimated values of the parameters are found to be unbiased and consistent. Next is the construction of the robust slope estimator using a 20% trimmed mean based on the nonparametric method. The robustness of this method is compared with the maximum likelihood method for replicated LFRM. Simulation results show that the 20% trimmed mean performs well even the datasets have a high number of outliers. The second part of the study focuses on outlier detection in replicated LFRM using COVRATIO statistic. The cut-off points and the performance of the method are obtained from the simulation study. From simulation results, the cut-off points obtained and power of performance is suggested that the COVRATIO statistic can be used to detect a single outlier in replicated LFRM. The last

part of the study concentrates on proposing a practical group method in clustering analysis. The motivation is to transform observation that are of unreplicated data to replicated data. Three clustering methods are considered and simulation studies are used to assess the performance of the parameter estimate of replicated LFRM. The benefits of these approach is that it can be done without making an assumption on the ratio of error variances. The applicability of all proposed methods is illustrated in published datasets.

Keywords: clustering, errors-in-variables model, mean square error, slope parameter

ABSTRAK

Kajian ini memfokuskan kepada penganggaran parameter terutama ketika kehadiran data terpencil, pengesanan data terpencil dan pengelompokan data dalam model linear hubungan fungsian. Terdapat dua kategori bagi model ini iaitu model tanpa replikasi dan model bereplikasi. Kajian ini dimulakan dengan pengubahsuaian kepada kebolehjadian maksimum untuk model linear hubungan fungsian tanpa replikasi apabila tiada maklumat mengenai nisbah ralat varians. Penganggar cerun teguh yang dinamakan kaedah kebolehjadian maksimum terubahsuai dicadangkan. Keputusan simulasi menunjukkan kaedah yang disyorkan adalah tidak dipengaruhi oleh data terpencil dan memberikan keputusan yang lebih baik daripada kaedah kebolehjadian maksimum. Kemudian, penambahbaikan penganggaran parameter dicadangkan dengan memperkenalkan data replikasi seimbang dalam model linear hubungan fungsian apabila tiada maklumat mengenai nisbah ralat varians. Anggaran parameter menggunakan kaedah kebolehjadian maksimum bersama dengan matriks asimptotik varians-kovarians menggunakan matriks maklumat Fisher diterbitkan. Keputusan daripada kajian simulasi menunjukkan nilai penganggar adalah saksama dan konsisten. Seterusnya, kaedah teguh tak berparameter dibangunkan menggunakan 20% min terpangkas. Perbandingan dilakukan di antara keteguhan kaedah ini dengan kaedah kebolehjadian maksimum bagi model linear hubungan fungsian bereplikasi. Keputusan menunjukkan kaedah 20% min terpangkas adalah baik walaupun terdapat banyak data terpencil. Bahagian kedua kajian memfokuskan kepada pengesanan data terpencil dalam model linear hubungan fungsian bereplikasi menggunakan statistik COVRATIO. Titik potongan dan kuasa pretasi diperolehi daripada kajian simulasi. Berdasarkan keputusan simulasi, titik potongan diperolehi dan dicadangkan bahawa kuasa prestasi statistik COVRATIO dapat mengesan satu nilai data terpencil dalam model linear hubungan fungsian bereplikasi. Bahagian

terakhir kajian memberi tumpuan kepada pembahagian data secara praktikal menggunakan analisis berkelompok. Tujuannya adalah untuk menukarkan data tanpa replikasi kepada data bereplikasi. Tiga kaedah berkelompok dipertimbangkan dan kajian simulasi dijalankan untuk menilai kuasa prestasi bagi anggaran parameter model linear hubungan fungsian bereplikasi. Kelebihan pendekatan ini ialah kaedah ini boleh dilakukan tanpa membuat sebarang andaian mengenai nisbah ralat varians. Penggunaan kesemua kaedah yang dicadangkan ditunjukkan dengan menggunakan contoh data set yang telah diterbitkan.

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful.

Alhamdulillah. All praises to Allah the Most Graceful and Most Merciful for giving me the strength and opportunity to complete this thesis. First and foremost, I am extremely grateful to my dedicated supervisor Prof. Dr. Yong Zulina Zubairi and my respectable supervisor Prof. Dr. Abdul Ghapor Hussin for their invaluable advice, continuous support and patience during my Ph.D. study. Their immense knowledge and guidance helped me in this research and complete this thesis. I would like to extend my sincere thanks to Universiti Malaya (UM), Universiti Pertahanan Nasional Malaysia (UPNM) and Kementerian Pengajian Tinggi (KPT) for financial support me to pursue this journey.

Special thanks to my beloved husband Nik Mohd Khashimi bin Hanafi and my lovely children Nik Muhammad Haiqal, Nik Emilyn and Nik Eryna who always give love, unparalleled support, motivation and patience in surviving this journey. Without their love and affection support that I received made me strong to overcome the difficulties in accomplishing my studies. I also dedicated special thanks to my beloved late father and mother who passed away during my Ph.D. journey. Although both of you are not here with me, I am very grateful for all sacrifices that you had done. I would like also to thank my dearest siblings and friends for giving me moral support to finish this thesis. Thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

X	:	Mathematical variable for a functional relationship model that is linearly related with Y
Y	:	Mathematical variable for a functional relationship model that is linearly related with X
x	:	Independent variable
y	:	Dependent variable
α	:	Intercept parameter
β	:	Slope parameter
δ_i / δ_{ij}	:	Error term for the independent variable
ε_i / ε_{ij}	:	Error term for the dependent variable
λ	:	Ratio of the error variances parameters in a functional relationship model
p	:	Number of group
m	:	Number of elements in a group

LIST OF ABBREVIATIONS

COVRATIO	:	Covariance Ratio
DIFFITS	:	Difference in fits
DFBETA	:	Difference in Beta
EB	:	Estimated Bias
EIVM	:	Errors-in-variable models
LFRM	:	Linear Functional Relationship Model
MLE	:	Maximum likelihood estimation
MMLE	:	Modified maximum likelihood estimation
MSE	:	Mean square error
SD	:	Standard deviation

LIST OF APPENDICES

CHAPTER 1: RESEARCH FRAMEWORK

1.1 Background of the Study

A study of the relationship between variables is a well-researched topic. One simple example is the relationship between two variables can be described as the fitting of straight lines. The most common method to model the relationship between two variables, namely the dependent variable and independent variable, is the linear regression model. The standard linear regression model assumes that the independent variables involved are measured without error. In contrast, the errors-in-variables model (EIVM) is a regression model that takes into account the measurement errors in the independent variables (Koul & Song, 2008). Errors could appear in experimental or in individual variability when the goal is to estimate the relationships between groups or population. Also, if errors in the dependent variables are ignored, the estimators obtained by classical regression are biased and inconsistent (Buonaccorsi, 2010).

Errors-in-variables model (EIVM) or measurement error model (MEM) was first introduces by Adcock in 1878 when he wanted to fit a straight line to bivariate data when both variables are measured with error. He proposed fitting a straight line can be done by minimizing the sum of the squares of the perpendicular distances of the points from fitted lines when both variables are subjected with errors and had equal variances. Over the years, the study of errors-in-variables model has gained importance and drawn a lot of attention among statisticians (Lindley, 1947; Madansky, 1959; Kendall & Stuart, 1979; Anderson, 1984; Fuller, 1987; Gillard, 2007).

Practical applications of errors-in-variables model are observed in almost every discipline such as in biology, ecology, economics, environmental sciences, manufacturing and others; for example, in environmental sciences, measuring the level of household lead is an error-prone process, not only because of device error but also the lead levels are exposed to many other factors such as air, dust, and soil with possibly correlated errors (Carroll, 1998). In manufacturing, for example, where sorting the manufactured goods to achieve certain standard or tolerance could be quite expensive. Thus, the most faster and cheaper test are also subjected to inspection errors (Buonaccorsi, 2010). An example in epidemiological studies is on diagnostic procedure in a blood test or an imaging technique where measurement error will lead to false result in a disease status (Buonaccorsi, 2010). Another example is in measuring nutrient intake, measurement error could occur in food frequency questionnaires and also in nutrient instruments used such as the food records (Carroll, 1998). The cause of coronary heart disease due to systolic blood pressure is another example in measurement error that occur on observed variability (Carroll, 1998). Most of the variables in these disciplines cannot be recorded correctly as mentioned in the examples above. Consequently, ignorance of measurement errors directly affects the desirable criteria of an estimator.

Over the past fifty years, many researchers have been working on the problem of estimating the parameters in the linear functional relationship model (LFRM), a subtopic in the errors-in-variables model. This model evaluates the relationship between the variables, both measured with error. From this point of view, the linear functional relationship model is more appropriate than the common linear regression model. Linear functional relationship model can be divided to unreplicated and replicated linear functional relationship model with certain recommendations (Dorff & Gurland, 1961a). The focus of this study is to examine methods of estimating the parameters of these two types of models. This includes obtaining estimations when outliers are present. Outliers

are observations that lies abnormal distance with the remainder of other observations. When estimating parameters for most models, it is worthwhile to note that most of the methods in the literature are heavily reliant on the normality assumption. However, when outliers are present in the data set, these can lead to errors in parameter estimation. To overcome the situation when the outliers exist in the dataset, a robust method is needed to estimate the parameter in linear functional relationship model. A robust method is referring to the ability of the parameter estimates to remain unaffected even in the presence of a single outlier. Thus, in this study, a robust method using the modified maximum likelihood method is proposed to estimate the slope parameter in unreplicated linear functional relationship model.

When estimating parameters in the unreplicated linear functional relationship model, the ratio of the error variances must be known to obtain the parameter estimate (Lindley, 1947). However, this is not necessary for a replicated model. The error variances may be easily estimated without having the assumption on the ratio of the error variances parameters if the replication of the observations is available (Barnett, 1970). The focus of this study is the parameter estimation and the covariance matrix of the replicated linear functional relationship model for the case balanced observations.

As mentioned before, the methods available in estimating the parameters are based on the normality assumption including for replicated linear functional relationship model. To overcome the presence of outliers in the dataset, the robust approach that is considered is the nonparametric method. The proposed method is an extension of the idea from Al-Nasser and Ebrahem (2005) and Ghapor et al. (2015) where the estimate of the slope parameter for replicated linear functional relationship model is obtained.

Another area of the research is a handling the presence of outliers in the datasets. An outlier is a point or some points of observation that not follow the pattern of the rest of

the observations. Past studies have considered different diagnostic tools to detect outliers in the linear functional relationship model; for example, Abdullah (1995) applied diagnostic methods in regression analysis to the functional model while Nurunnabi et al. (2011) proposed group deleted version to identify outliers. Moreover, Ghapor et al. (2014) proposed *COVRATIO* procedure in detecting a single outlier in unreplicated linear functional relationship model. In this study, the idea of *COVRATIO* statistic, which has been used in unreplicated model is extended to replicated linear functional relationship model.

The third area in this study is on grouping the data from unreplicated linear functional relationship model into a replicated linear functional relationship model using clustering method. This approach may provide a solution when information about ratio of error variances from the unreplicated linear functional relationship model is not available. In other words, the idea of clustering can be extended to create groupings and thus can be used to estimate the parameters using the replicated linear functional relationship model.

1.2 Problem Statement

Generally, this study addresses three problems in linear functional relationship model (LFRM). The first problem is related to unidentifiability problem in estimating the parameters of linear functional relationship model. The parameters in linear functional relationship model are the intercept α, the slope β, two error variances, σ^2 and τ^2 and also the incidental parameters, X_i. Unidentifiability occurs when the number of parameters increase in proportion to the number of observations, i. This will lead inconsistencies in linear functional relationship model due to the presence of the

incidental parameter. Although numerous studies in parameter estimation have been done in the past (Lindley, 1947; Kendall, 1952; Villegas, 1961; Moran, 1971; Wong, 1989), little attention has been paid to obtain robust slope parameter estimator in linear functional relationship model. To overcome this problem, a robust estimator for the slope parameter of the unreplicated linear functional relationship model is needed especially in the presence of outliers. The replicated model is known to avoid the unidentifiability problem. This is because, in the replicated model, as the number of sample sizes increases, the number of parameters remains unchanged. As a result, it is necessary to improve the method of parameter estimation for replicated linear functional relationship model so that all the parameters can be estimated. Here, a robust nonparametric method to estimate the slope parameter in replicated linear functional relationship model is considered in the presence of the outliers.

The second part of this study is related to the presence of outliers that are often unavoidable. Before making an analysis, it is important to detect outliers as their presence in the datasets give an adverse impact on the statistical analysis. Outliers problems in a linear regression model and circular regression model have been widely addressed (Belsley et al., 1980; Rousseeuw & Leroy, 1987; Maronna et al., 2006; Ibrahim et al., 2013). In contrast, methods for detecting outliers in replicated linear functional relationship model have never been investigated. Finding an appropriate method for detecting outliers in a balanced replicated linear functional relationship model has become an inevitable requirement.

The third area in this study is addressing the unidentifiability problem in the linear functional relationship model. It needs an assumption or information on the ratio of error variances to estimate the parameters. However, this information is not available (Klepper & Leamer, 1984). It is worthy to explore if this problem is present in the replicated linear

functional relationship model. It is important to overcome this problem so that one can find the estimation of all parameters, in particular, the slope parameter.

1.3 Objectives of Research

The main objective of this study is to propose methods of parameter estimation and a method to detect and identify outliers in linear functional relationship models. The specific objectives are given as follows:

1. to propose a modified maximum likelihood estimation for the slope parameter in unreplicated linear functional relationship model.
2. to derive the parameter estimator as well as variance-covariance matrix for balanced replicated linear functional relationship model.
3. to develop a new technique using nonparametric method to estimate the slope parameter in balanced replicated linear functional relationship model.
4. to identify the outliers by using the COVRATIO statistic in balanced replicated linear functional relationship model.
5. to propose a new grouping approach using clustering analysis in linear functional relationship model.

1.4 Methodology and Flow Chart of Study

In the first part of this study, a detailed literature review is done on the historical background, current issues and problems arising on errors-in-variable model, linear

functional relationship model, nonparametric methods, outliers, and clustering analysis topics.

From the literature review, a robust method is proposed by modifying the maximum likelihood estimation method by assuming the ratio of error variances is equal to estimate the slope parameter in unreplicated linear functional relationship model. The performance of this method is compared with the existing maximum likelihood estimation method through the simulation study with performance measures of estimated bias and mean square error are used. Additionally, two data sets are used to illustrate the practical application of this method.

Next, is the parameter estimation for replicated linear functional relationship model. The maximum likelihood estimation is considered to estimate the parameters of replicated linear functional relationship model for data with balanced observations in each group. The estimation of parameters as well as the covariance matrix for the estimated parameters are obtained using the Fisher information matrix. A general guideline to group the observations is proposed and presented. The performance of the estimated parameters is obtained via simulation study using measures of the estimated bias, the mean square error and the standard deviation respectively. The applicability of this model is illustrated using two data sets.

Then, a robust nonparametric estimation is developed to estimate the slope parameter for the replicated linear functional relationship model which is based on balanced observations in every group. Through the simulation study, the robustness of this method is compared with the existing maximum likelihood estimation method using measures of estimated bias and mean square error. A simulation study is performed and two data sets are used to illustrate the application of this method.

This is followed by an outlier detection method for the model considered. For outlier detection in replicated linear functional relationship model which is based on balanced observations in every group, the *COVRATIO* statistic is considered. A simulation study is performed to find the cut-off point. After finding the cut-off point, the power of the performance is investigated also through the simulation study. The applicability of the proposed method is illustrated using two data sets namely a simulated data set and a real data set.

Finally, the clustering technique is considered to group the observations from unreplicated data and then an estimation of the slope parameter is obtained using replicated linear functional relationship model. Through a simulation study, the estimation of the slope using three different clustering techniques is compared with the baseline slope estimation from unreplicated linear functional relationship model by mean square error. Also, the applications using a simulated data set and a real dataset are illustrated.

All methods of estimation and outlier detection mentioned above are developed and conducted using R programming. R is a free software and open-source. As mentioned above, all proposed methods will be validated using simulation study and will be illustrated using data sets. Figure 1.1 shows the flow chart of this study.

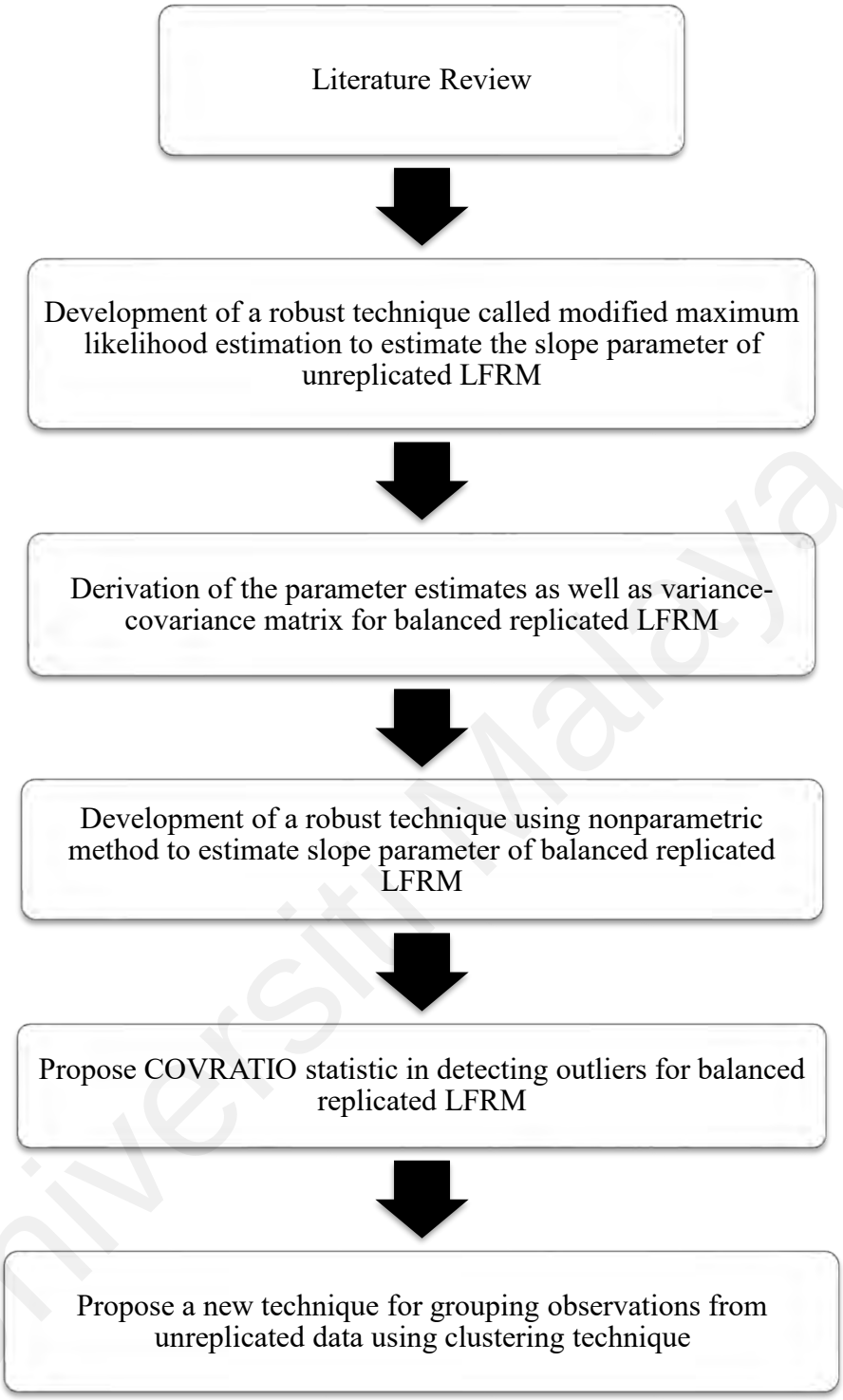

Figure 1.1: Flow chart of the study

1.5 Source of Data

In this study, the following data for illustration and application are used. Full data sets are given in APPENDICES (APPENDIX A - APPENDIX D). The following are the background of the data sets used in this study.

1) Fat Measurement Data from Goran et al. (1996)

The purpose of this study was to examine the accuracy of some widely used body-composition techniques for children through the use of the dual-energy X-ray absorptiometry (DXA) technique. Subjects were children between the ages of 4 and 10 years. The fat mass measurements taken on the children are by using two techniques; skinfold thickness (ST) and bioelectrical resistance (BR). This data set is used in Chapter 4 and Chapter 7.

2) Frosted Flakes data from Maindonald and Braun (2010)

The purpose of this study was to measure the sugar concentrations (in percentage) for approximately 25 g of cereal samples measured by two techniques, namely the high-performance liquid chromatography (a slow and accurate laboratory method) and a quick method using the infra-analyser 400 (IA400). This data set is used in Chapter 3.

3) Systolic Blood Pressure from Bland and Altman (1999)

The purpose of this study was to measure the systolic blood pressure which simultaneous measurements were made by two experienced observers denoted as J (x_{ij}) and R (y_{ij}). The data set is taken from Bland and Altman (1999). This data set is used in Chapter 4 and Chapter 6.

4) Iron in Slag Data from Hand et al. (1994)

The purpose of this study is to measure the iron content of crushed blast-furnace slag measured by two different techniques, which are chemical test and magnetic test. The data set has 50 observations results of iron content which taken from Hand et al. (1994). This data set is used in Chapter 5.

1.6 Thesis Organization

This thesis is divided into seven chapters. Chapter One briefly introduces the research framework which includes the background of errors-in-variables model, followed by problem statement, objectives of the study, methodology and the flowchart of the study, the source of data and the thesis organization. Chapter Two reviews the literature and historical background of the errors-in-variables model. These includes the parameter estimation of unreplicated linear functional relationship model and replicated linear functional relationship model. Furthermore, discussions on outliers, robust statistics and clustering analysis are given. Chapter Three presents the modification of maximum likelihood method for unreplicated linear functional relationship model assuming the error variance ratio equal one. Chapter Four discusses on maximum likelihood estimation to estimate the parameters of replicated linear functional relationship model which is based on balanced observations in every group. Chapter Five proposes a robust nonparametric method to estimate the slope parameter in balanced replicated linear functional relationship model which is based on non-normality assumptions. Chapter Six proposes a *COVRATIO* statistic to detect outliers in the balanced replicated linear

functional relationship model. Chapter Seven presents a new technique to estimate the slope parameter of replicated linear functional relationship model assuming the observations can be grouped from unreplicated linear functional relationship model using clustering method. Chapter Eight presents the summary of the study, contribution and suggestions for future works. Also, the list of references used in this study is given along with the list of publications and oral presentations. Finally, the list of appendices to support this work is given at the end of thesis.

CHAPTER 2: LITERATURE REVIEWS

2.1 Introduction

This chapter presents literature review that leads to this study. Some fundamental concepts are presented related to the development of this study. A brief review on errors-in-variable model (EIVM) is given in Section 2.2 followed by linear functional relationship model (LFRM) in Section 2.3 including the unreplicated model and replicated model. Section 2.4 reviews the background information on the topics of outliers and robust statistics. Lastly, Section 2.5 review on clustering analysis including the similarity measures and agglomerative hierarchical clustering.

2.2 Errors-in-variables Model

Errors-in-variables model has been introduced by Adcock (1878) and become an important topic since a century ago. Adcock used the least squares method for the estimation of the slope parameter by assuming both variables have equal error variance when he investigated the estimation properties in ordinary linear regression models when both variables namely independent variable, x and dependent variable, y are measured with errors. However, the study was quite restricted and only focused on equal error variances and the method has been known as orthogonal regression. Orthogonal regression minimizes the orthogonal distances from the data points to the regression line.

A year later, in 1879, Kummel extended Adcock's study by assuming the ratio of error variance is known, but not necessarily equal to one. However, Kummel argued that the knowledge of the ratio of error variances is hardly known and the practitioners should have sufficient knowledge about the ratio. Extending the work of Adcock, Pearson (1901) showed that the slope parameter of orthogonal regression lies between the regression line of y on x and x on y. Not until a few years later, the idea of orthogonal regression was introduced by Deming's (1931) and this method is often referred to as Deming's (1931) regression. He showed that his method has considered unequal error variances.

A different method was proposed by Wald and Wolfowitz (1940), in estimating the slope parameter by dividing the order of explanatory variables into two groups without taking any assumption about the error structure. He obtained the slope estimator by dividing the observations into two equal groups and finding the group means. Later on, Bartlett (1949) extended Wald's idea by distributing the order of explanatory variables into three groups to get a more efficient estimator for the slope. Another grouping method of the explanatory variables based on some specific assumptions was suggested by Neyman and Scott (1951). Review on grouping methods has been discussed by Madansky (1959). Furthermore, Dorff and Gurland (1961b) compared different consistent slope estimators on the basis of asymptotic variances in which they obtained from the method of grouping.

Another method in estimating the parameters errors-in-variables model is the geometric mean method. This method gives the minimum sum of products of the horizontal and vertical distances of the observations from the lines in estimating the slope parameter (Teissier, 1948). Geometric mean method had been used extensively in fisheries. Further studies on geometric mean method can be found in Jolicoeur (1975) (1975) and Barker, Soh, and Evans (1988).

Numerous studies have used the method of moments to estimate parameters in errors-in-variables model. Method of moments is one of the superior method if the estimation of parameters comes from a known family of probability distribution. In certain cases, the estimators from method of moments can be calculated easily compared with the other methods. In 1949, Geary introduced the method of moments in errors-in-variables model but using cumulants. He also discussed the method of moments for large sample sizes from his earlier work (Geary, 1942). Then, in 1951, Drion computed the variance of the sample moments and showed that his slope estimate is consistent. Later on, Pal (1980) and Montfort (1989) have considered this method in order to achieve the optimal estimators using estimators based on higher moment. Recently, Dunn (2004) derived the formulas for the slope estimators using a method of moments based on the first and second moments. Later, Gillard (2014) described in details that the higher moments have larger variance and give some recommendations in estimating the slope parameters.

The use of maximum likelihood estimation methods in estimating parameters of errors-in-variable model was first introduced by Lindley (1947). The maximum likelihood estimation is a method that determines values for the parameters of a model by maximizing the likelihood of the model. Lindley (1947) showed that the likelihood equations are inconsistent and need some prior information on the parameters to solve this problem. He also mentioned that the most common assumption is the value of the ratio of error variances is known. Later, Bayesian approach in estimating the parameters in errors-in-variables model has been suggested by Lindley and El-Sayyad (1968) based on prior information. They concluded that the likelihood approach may be misleading in some ways. It is worthwhile to note others authors like Birch (1964), Barnett (1970) and Wong (1989) have considered the likelihood method to estimate the parameters in order to get the consistent estimation.

A different approach namely total least square method has been investigated. Total least square method minimizes the sum of the squared orthogonal distances from the data. Many authors have been explored total least-squares method in estimating the parameters of errors-in-variables model (Golub & Van Loan, 1980; Van Huffel & Vandewalle, 1991; Van Huffel & Lemmerling, 2002). More least square problems, solutions and applications have been discussed by Markovsky and Van Huffel (2007).

Applications involving errors-in-variables model can be seen in many fields. The method of least squares has been commonly used in computational mathematics and engineering optimization problem. Another application can be seen in wavelet filtering by Gençay and Gradojevic (2011). They indicated that this approach does not require instruments and yields unbiased estimates for the intercept and slope parameters although this approach still requires a lot more research. Also, O'Driscoll and Ramirez (2011) considered the geometric form of errors-in-variables model. They analysed the performance of various slope estimators including an adjusted fourth moment estimator proposed by Gillard (2014) to remove the jump discontinuity in the estimator of Copas (1972). Other authors, for instance, Doganaksoy and Van Meer (2015) had been used errors-in-variable model for assessing performance of the semiconductor devices. Nowadays, new technologies have vastly improved data collection resulting in an avalanche of data from various discipline such as the financial sectors. In the financial sector, data-driven decisions that accelerate innovation, improve customer experience and reduce costs are becoming increasingly popular and these would become variables where errors-in-variables problems would arise (Mirmozaffari et al. 2021; Sharif, et al. 2019). As mentioned by Mirmozaffari et al., (2020), using the errors-in-variable models is critical because errors can occur due to datasheet manipulation or the availability of some missing values in the data collection.

The errors-in-variables model is described as follows:

$$Y = \alpha + \beta X$$

where the observations of X and Y are linearly related and the parameter α and β is the intercept and slope parameter respectively. If both observations on X and Y can be observed exactly, the value of α and β can be solved using simultaneous linear equation. If only the variable Y is observed with error, the parameters α and β can be solved using ordinary linear regression. However, if observations on both X and Y are measured with error, then the parameters of α and β need to be solved using errors-in-variables model. There are some differences between linear regression model and errors-in-variables model. For example, the variables X and Y are symmetric in errors-in-variables model; unlike in linear regression model, the variable X and Y is called the independent and dependent variables respectively. Another difference is that in errors-in-variables model, it allows sampling variability or errors of both variables X and Y. Furthermore, errors-in-variables model can also be considered as an extension of the linear regression model. In addition, in errors-in-variables model, there is no distinction between independent and dependent variable for X and Y. In reality, these two variables X and Y cannot be observed directly as their measurements are subject to error.

As mentioned by Kendall and Stuart (1979), there are three models under the errors-in-variables model by considering the values of X_i as follows:

i) Functional relationship model where X_i is a mathematical variable or fixed constant μ.

ii) Structural relationship model where X_i is a random variable with mean μ and variance σ_X^2.

iii) Ultrastructural relationship model where there is a combination of the functional and structural relationship as introduced by Dolby (1976).

However, in practice, applications using data do not differentiate the functional or structural model (Sprent, 1990). This study will focus on the functional relationship model (LFRM) where the variable X is a mathematical variable and the data sets are linear. The linear functional relationship model will be discussed in detail in the following section.

2.3 Linear Functional Relationship Model

In this section, the focus is on linear functional relationship model which the underlying variable X is a fixed or deterministic. The variable X can be defined as a constant and mathematical variable without specific distributional properties (Kendall, 1952; Fah, Hussin, & Rijal, 2007). Over the past few decades, numerous authors have been working on functional model, a subtopic in errors-in-variables model (Lindley, 1947; Kendall & Stuart, 1979; Wong, 1989; Gillard & Iles, 2006). Many authors used the maximum likelihood estimation method for parameter estimation with the assumption that the dependent and independent variables are joint normally and identically distributed.

Let X_i and Y_i are two mathematical variables that are linearly related as follows:

$$Y_i = \alpha + \beta X_i \tag{2.1}$$

where α is the intercept and β is the slope parameters. Assume that for each x_i and y_i are measured with errors δ_i and ε_i instead of X_i and Y_i respectively and $i = 1,2, \dots, n$. Then it can be modelled as,

$$x_i = X_i + \delta_i \text{ and } y_i = Y_i + \varepsilon_i \tag{2.2}$$

where the error terms δ_i and ε_i are assumed to be mutually independent and normally distributed random variables,

$$\delta_i \sim N(0, \sigma^2) \text{ and } \varepsilon_i \sim N(0, \tau^2) \tag{2.3}$$

The above equation (2.3) show that the variances of the error term are not dependent on i and thus independent of the level of X and Y. This implies that

i) both errors have mean 0, that is $E(\delta_i) = 0$ and $E(\varepsilon_i) = 0$ where $i = 1,2, \dots, n$

ii) both errors have constant but different variance, that is $Var(\delta_i) = \sigma^2$ and $Var(\varepsilon_i) = \tau^2$ where $i = 1,2, \dots, n$.

iii) the errors are uncorrelated, that is $Cov(\delta_i, \delta_j) = 0$ and $Cov(\varepsilon_{i,} \varepsilon_j) = 0$, where $i \neq j$; $i, j = 1,2, \dots, n$ and $Cov(\delta_i, \varepsilon_{i,}) = 0$ where $i, j = 1,2, \dots, n$.

As given in equation (2.1), (2.2) and (2.3), there are $(n + 4)$ parameters which are the intercept, α, the slope, β, the error variances namely, σ^2 and τ^2 and also the incidental parameters, $X_1, X_2, \ldots, X_n$ (a vector of nuisance parameters which its dimensions may increase depends on the sample size) although those of the primary interest is β. For this type of model, one of the problem arise when the number of observations increase, it makes the number of parameters will also increase. In this case, the likelihood function is unbounded when there is only a single observation at each point. In order to solve this problem, some constraint or condition must be introduced or replicated data needs to be acquired.

Lindley (1947) first used the maximum likelihood estimation to estimate the parameter, he proposed that the ratio of two error variances need to be known in order to overcome the inconsistencies in the equation. Later, Moberg and Sundberg (1978) suggested that certain constraints by making assumptions on the variances and covariances of the errors to find the maximum likelihood estimation of parameters in a linear functional relationship model with normally distributed errors. The constraints include:

i) $Var(\delta_i)$, $Var(\varepsilon_i)$ and $Cov(\delta_i, \varepsilon_i)$ are all known.

ii) $\frac{Var(\varepsilon_i)}{Var(\delta_i)} = \lambda$ is known and $Cov(\delta_i, \varepsilon_i) = 0$.

Another possible solution to overcome the inconsistencies is to obtain replication of observations to acquire consistent estimate of parameters in particular for the slope estimate β (Klepper & Leamer, 1984). The scope of this research is on the estimation of the slope parameter, β, although other parameters may also have important role in linear functional relationship model. It is worthwhile to note that as mentioned by Tony Cai and Hall (2006), the majority attention usually focuses in estimating the slope parameter, β

compared to the intercept parameter, α as from a theoretical point of view, the role of α is minor. For this study, two situations will be considered, namely, when the ratio of two error variances namely λ is known and also when replicated observations are available. Dorff and Gurland (1961a) extended the linear functional relationship model into two categories which are the unreplicated linear functional relationship model and replicated linear functional relationship model. Thus, by considering two situations mentioned before, the unreplicated linear functional relationship model is when the ratio of two error variances namely λ is known and replicated linear functional relationship model is when replicated observations are available. A detailed explanation for both unreplicated and replicated are given in subsequent section.

2.3.1 Unreplicated Linear Functional Relationship Model

Maximum likelihood estimation method is the most common method used for parameter estimation in linear functional relationship model. The maximum likelihood estimation method has certain optimal properties and it is possible to work out the asymptotic variance-covariance matrix of the estimators. Consider the log likelihood function of the linear functional relationship model below:

$$\log L\left(\alpha, \beta, \sigma^2, \tau^2, X_1, \dots, X_n; x_1, \dots, x_{n,}, y, \dots, y_{n,}\right) = -n\, log(2\pi) \tag{2.4}$$

$$-\frac{n}{2} log(\sigma^2) - \frac{n}{2} log(\tau^2) - \frac{\sum_{i=1}^{n}(x_i - X_i)^2}{2\sigma^2} - \frac{\sum_{i=1}^{n}(y_i - \alpha - \beta X_i)^2}{2\tau^2}$$

The likelihood function in (2.4) is unbounded by putting $\hat{X}_i = x_i$ and will approach infinity by considering σ^2 approaches to 0, irrespective of the values α, β and τ^2. Therefore, as mentioned by Lindley (1947), additional constraint is assumed which is $\tau^2 = \lambda\sigma^2$, where λ is known to avoid an unbounded problem in above equation. Thus, the log likelihood function becomes

$$\log L\left(\alpha, \beta, \sigma^2, X_1, \dots, X_n; \lambda, x_1, \dots, x_{n,}, y, \dots, y_{n,}\right) = -n\, log(2\pi) \tag{2.5}$$

$$-\frac{n}{2}\log\lambda - n\log(\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - X_i)^2}{2\sigma^2} - \frac{\sum_{i=1}^{n}(y_i - \alpha - \beta X_i)^2}{2\lambda\sigma^2}$$

From (2.5) above, there are $(n + 3)$ parameters to be estimated, which are the intercept α, the slope β, the error variance σ^2 and also the incidental parameters $X_1, X_2, \dots, X_n$. The parameters $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}^2$ and $\hat{X}_i$ can be obtained by differentiating $\log L$ with respect to parameters α, β, σ^2 and X_i. By setting the derivative to zero, the parameters are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{2.6}$$

$$\hat{\beta} = \frac{\left(S_y^2 - \lambda S_x^2\right) + \sqrt{\left(S_y^2 - \lambda S_x^2\right)^2 + 4\lambda S_{xy}^2}}{2S_{xy}}, \tag{2.7}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left\{\sum_{i=1}^{n}(x_i - X_i)^2 + \frac{1}{\lambda}\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}X_i\right)^2\right\}, \tag{2.8}$$

$$\text{and } \hat{X}_i = \frac{\lambda x_i + \hat{\beta}\ (y_i - \hat{\alpha})}{\lambda + \hat{\beta}^2}. \tag{2.9}$$

where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ are the sample means of data x and y respectively, S_x^2 and S_y^2 are the sample variances with the equation $S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$, $S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$ and $S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$ is the sample covariance.

As mentioned earlier, the interest in this study is getting the estimation of the slope parameter, β. Several methods have been considered in estimating the slope parameters in previous studies. Dent (1935) proposed the slope estimator called geometric mean functional relationship estimation which has been widely used in fisheries research as follows:

$$\hat{\beta} = Sign\big(Cov(x,y)\big)\left\{\frac{\sum(y_i - \bar{y})^2}{\sum(x_i - \bar{x})^2}\right\}^{\frac{1}{2}} \tag{2.10}$$

Later on, Sprent (1970) stated the estimator in (2.10) is not usually consistent. He suggested that although the underlying symmetry of x and y from the functional relationship model is still maintained, the estimator ignores the unidentifiability problem and also assumes normality without knowing the error variance.

Earlier Wald and Wolfowitz (1940) suggested a computation of arithmetic means by arranging the order of the value of x_i to find consistent estimator of β . These values are divided into two equal sub-groups namely the lower group of observations called $(\bar{x}_1, \bar{y}_1)$

and upper group of observations called $(\bar{x}_2, \bar{y}_2)$. This two-group method has the slope parameter estimated by

$$\hat{\beta} = \frac{(\bar{y}_2 - \bar{y}_1)}{(\bar{x}_2 - \bar{x}_1)} \tag{2.11}$$

In addition, Bartlett (1949) extended Wald and Wolfowitz (1940) method by dividing the ranked x_i into three equal groups. If the number of observations has a remainder when divided by three, then he will make it approximately equal. Only the lower $(\bar{x}_1, \bar{y}_1)$ and upper $(\bar{x}_3, \bar{y}_3)$ groups are taken to calculate the arithmetic mean while ignoring the middle group, and the slope estimated using three-group method as follows,

$$\hat{\beta} = \frac{(\bar{y}_3 - \bar{y}_1)}{(\bar{x}_3 - \bar{x}_1)} \tag{2.12}$$

However, as the upper and lower groups are not necessarily the same when ranked on y_i, both methods are not symmetric in x and y. Compared with the two-group method, the three-group method is more efficient as its variance does have the smallest possible values (Dorff & Gurland, 1961b). Nevertheless, both methods give the consistent estimate of β (Hussin, 2004). To overcome the asymmetrical problem, the arithmetic means also need to be calculated based on ranking on y_i.

Also, Housner and Brennan (1948) arranged x_i values in ascending order and the associated values of y_i which are not be in ascending order are taken. Although this

method gives a consistent estimate of β, the slope is not symmetric in x and y (Dorff & Gurland, 1961b; Hussin, 2004). The estimate of β for this method as follows:

$$\hat{\beta} = \frac{\sum i\left(y_{(i)} - \bar{y}\right)}{\sum i\left(x_{(i)} - \bar{x}\right)} \tag{2.13}$$

Later, Durbin (1954) proposed his ranking method by first ranking x's and y's in ascending order, on the basis of x values. Then, he ranked the x's and y's in ascending order, on the basis of y values. This method also gives a consistent estimate of β (Hussin, 2004). The slope estimates using Durbin's ranking method is given by

$$\hat{\beta} = \frac{\sum\left(x_{(i)} - \bar{x}\right)^2\left(y_{(i)} - \bar{y}\right)}{\sum\left(x_{(i)} - \bar{x}\right)^3} \tag{2.14}$$

Later, the modified least squares method proposed by Cheng and Ness (1994) has the assumption that the variance ratio $\lambda = \frac{\sigma^2}{\tau^2}$ is known. The estimation using the proposed method is the same as the same estimates in the maximum likelihood estimation method, but does not require the normality assumption. The slope estimate is given by

$$\hat{\beta} = \frac{\left(S_{yy} - \lambda S_{xx}\right) + \sqrt{\left(\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S_{xy}^2\right)}}{2S_{xy}} \tag{2.15}$$

where $S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{yy} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ respectively.

Also, Al-Nasser and Ebrahem (2005) introduced a nonparametric method which also does not require a normality assumption. According to Sprent and Smeeton (2016), there are several advantages by using the nonparametric method such as no prior knowledge needed on the distribution of the model and also the method is robust in the presence of outliers in the dataset. In Al-Nasser and Ebrahem (2005) method, they suggested by arranging x_i values in ascending order and the related values of y_i which are not be in ascending order are taken. In getting the slope parameter, he then finds the median of all slopes after listing down all the possible paired of slopes.

Later, Ghapor et al. (2015) extended Al-Nasser and Ebrahem (2005) method by adding one step which is also arranging y_i values in ascending order and the related values of x_i which are not be in ascending order are taken. The same procedure with the additional step for getting a slope parameter is applied as the Al-Nasser and Ebrahem method which is also using the median.

2.3.2 Replicated Linear Functional Relationship Model

In unreplicated linear functional relationship model, all the parameters cannot be estimated consistently. This is due to the fact that when the number of observations increases, the number of parameters will also increase. Unless some additional information is available such as the ratio of the error variances is known. However, as mentioned by Klepper and Leamer (1984) the additional information is either not

available or is not widely shared by researchers in the field and such problem can be avoided if this information can be derived from the sample itself. Replicated observations can be used to get the estimate error variances of the linear functional relationship model on the true values of the study and the explanatory variables are available. Furthermore, replication of observations allows consistent estimation when the ratio of error variances is unknown. This is due to the fact that when the number of observations increase, the number of parameters is still fixed and not increase.

Works on the topic of replication for different model have been discussed in the literature. The most detailed study by Dorff and Gurland (1961a) who compared various consistent slope estimators in terms of their asymptotic variances for both unreplicated and replicated model. The estimators are derived from the method of grouping and share properties with other estimators used in the unreplicated case. As mentioned by Kendall and Stuart (1979), the unidentifiability problem of the unreplicated model can be overcome by using some properties of replicated errors. Dolby (1976) derived the maximum likelihood estimators of the slope parameter by synthesizing linear functional and linear structural model called the ultrastructural model. Further development on the ultrastructural relationship model can be seen in Shalabh, Paudel, and Kumar (2009) where they proposed consistent estimation of parameters in replicated ultrastructural relationship model. They also derived the asymptotic efficiency properties of the estimators. Singh, Jain, and Sharma (2012) proposed replicated ultrastructural consistent estimators by assuming some prior information available regarding regression coefficients in the form of stochastic linear restriction and in 2014, in other form called exact linear restrictions (Singh, Jain, & Sharma, 2014).

In linear structural relationship model, Chan and Mak (1979) derived the maximum likelihood estimators of the slope parameter and showed that the estimator is consistent

when the number of replicates increase. Isogawa (1985b) has discussed on the replicated case in multivariate linear structural relationship model and derived the asymptotic covariance matrix. In the same year, she also investigated and concluded that the generalized least-squares method is asymptotically efficient with replicated observations (Isogawa, 1985a). She then continued her works on linear structural relationship model for unpaired and unequally replicated observation by considering five models with different types of the error variances (Isogawa, 1992).

The early work on replicated linear functional relationship model can be found on Barnett (1970). The maximum likelihood estimation in replicated linear functional relationship model was derived when considering the alternate models for different error structures that might be applicable in biological and medical field of research. However, he mentioned that there was no closed form can be found and iterative solution might be needed. Although he gave few ideas to start the iteration process, he also mentioned that the replicated linear functional relationship model can be explored in terms of iterative procedure and initial point estimate. Further works can be found for nonlinear replicated linear functional relationship model (Dolby and Lipton 1972) and for unpaired and unequally replicated data in linear functional relationship model by giving some recommendations on five different models depending on the error variances (Dolby et al. 1987). Hussin (2005) and Mokhtar et al. (2017) have discussed the parameter estimation for the replicated linear functional relationship model on circular variables.

Assume that X_i and Y_i are two random variables, there may be replicated observations of X_i and Y_i occurring in p groups. A linear relationship between X_i and Y_i are given by

$$Y_i = \alpha + \beta X_i \tag{2.16}$$

$$\text{where } x_{ij} = X_i + \delta_{ij} \text{ and } y_{ik} = Y_i + \varepsilon_{ik} \tag{2.17}$$

for $i = 1,2,\cdots,p$, $j = 1,2,\cdots,m_i$ and $k = 1,2,\cdots,n_i$. The errors terms are δ_{ij} and ε_{ik} follow normal distribution with mean zero and variance σ^2 and τ^2 respectively i.e. $\delta_{ij} \sim N(0,\sigma^2)$ and $\varepsilon_{ik} \sim N(0,\tau^2)$. This implies that

i) both errors have mean 0, that is $E(\delta_{ij}) = 0$ and $E(\varepsilon_{ik}) = 0$ where $i = 1,2,\ldots,p$ and $j = 1,2,\cdots,m_i$ and $k = 1,2,\cdots,n_i$.

ii) both errors have constant but different variance, that is $Var(\delta_{ij}) = \sigma^2$ and $Var(\varepsilon_{ik}) = \tau^2$ where $i = 1,2,\ldots,p$ and $j = 1,2,\cdots,m_i$ and $k = 1,2,\cdots,n_i$.

The estimation of parameter may be obtained by maximum likelihood estimation which involves some iterative procedures. In this case, the log likelihood function can be expressed as

$$\log L\left(\alpha,\beta,\sigma^2,\tau^2,X_1,\ldots,X_p;x_{ij},y_{ik}\right) = -\frac{1}{2}\left(\sum_{i=1}^{p} m_i + \sum_{i=1}^{p} n_i\right)\log 2\pi \tag{2.18}$$

$$-\frac{1}{2}\left(\sum_{i=1}^{p} m_i \log\sigma^2 + \sum_{i=1}^{p} n_i \log\tau^2\right)$$

$$-\frac{1}{2}\left\{\sum_{i=1}^{p}\sum_{j=1}^{m_i}\frac{\left(x_{ij}-X_i\right)^2}{\sigma^2} + \sum_{i=1}^{p}\sum_{k=1}^{n_i}\frac{(y_{ik}-\alpha-\beta X_i)^2}{\tau^2}\right\}$$

There are $(p+4)$ parameters to be estimated which are the intercept α, the slope β, the error variances σ^2 and τ^2 and also the incidental parameters $X_1, X_2, \ldots, X_p$. These parameters may be obtained by differentiating the log likelihood function as given in (2.18) with respect to $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2$ and $\hat{X}_i$ respectively and equating to zero. As mentioned by Barnett (1970), the closed form solution for these parameters are not available. Thus, to overcome this, the iterative method is used instead. Thus, the parameters can be obtained in order given by

$$\hat{X}_i = \frac{1}{\hat{\Delta}}\left\{\frac{m_i \bar{x}_{i.}}{\hat{\sigma}^2} + \frac{n_i \hat{\beta}}{\hat{\tau}^2}(\bar{y}_{i.} - \hat{\alpha})\right\}, \tag{2.19}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{m_i}\left(x_{ij} - \hat{X}_i\right)^2}{\sum_{i=1}^{p} m_i}, \tag{2.20}$$

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{p}\sum_{k=1}^{n_i}\left(y_{ik} - \hat{\alpha} - \hat{\beta}\hat{X}_i\right)^2}{\sum_{i=1}^{p} n_i}, \tag{2.21}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{p} n_i\left(\bar{y}_{i.} - \hat{\beta}\hat{X}_i\right)}{\sum_{i=1}^{p} n_i}, \tag{2.22}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{p} n_i \hat{X}_i(\bar{y}_{i.} - \hat{\alpha})}{\sum_{i=1}^{p} n_i \hat{X}_i^2} \tag{2.23}$$

where $\bar{x}_{i.} = \frac{\sum_{j=1}^{m_i} x_{ij}}{m_i}$, $\bar{y}_{i.} = \frac{\sum_{k=1}^{n_i} y_{ik}}{n_i}$ and $\hat{\Delta}_i = \frac{m_i}{\hat{\sigma}^2} + \frac{n_i \hat{\beta}^2}{\hat{\tau}^2}$.

Note that $\bar{x}_{i.}$ and $\bar{y}_{i.}$ are the sample means in each group. However, the iteration procedures that have been described by Barnett (1970) are not very ideal. Thus, an improvement to this model is needed to find the values of the estimated parameters.

2.4 Outliers and Robust Statistics

In this section, a briefly explanation about two situations in the data set which give a major impact in the analysis. The first situation is the existence of outliers and the second is the procedure or method called robust statistics that cater the outliers in the data set.

2.4.1 Outliers

The observation is considered as an outlier if it does not follow any pattern with the remainder of the other observations. The possibility of these observations become outliers need to be checked thoroughly. Outliers become a common problem and may occur as the results of misplaced decimal points, mistakes in keypunch and mistakes in sampling the population. Outlier can happen when the observations in an experiment is incorrectly observed or recorded or data set is mistakenly entered in the computer (Cateni et al., 2008). The issue of outliers has received considerable critical attention among researchers. Outliers should be identified on the basis of their potential effect on the estimation of the parameters, the precision of the estimated parameters and also the overall predictive capacity of the model. As quoted by Hampel et al., (2011), “A routine

data set typically contains about 1-10% outliers and even the highest quality data set cannot be guaranteed free of outliers".

Numerous authors have considered outliers in linear model such as Montgomery, Peck, and Vinning (2012), Hussin et al. (2013) and Satman (2013). In linear regression model, there are different types of outliers which depends on the location of one observation or a few observations that give an impact on the model. First, the y-outlier or widely known as influential observation happen if the parameter estimates change significantly when a point is removed from the calculation. Next, is the x-outlier or the leverage point which occur if the point has a greater ability to move the regression line. Residual outlier on the other hand happens when a point has a large standardized (deletion) residual. These outliers may result in substantial changes to the parameter estimates and prediction of the model. As mentioned by Chatterjee and Hadi (1988), there is interrelationship among influential observation with high leverage points and residual outliers.

Several outlier diagnostics are available in the literature for linear regression including Cook's distance, Difference in fits (DIFFITS), Difference in Beta (DFBETA), Covariance Ratio (*COVRATIO*) and others. Most of the ideas of finding influential observations in regression are developed on the basis of deleting the observations one after another and measuring their effects on various aspects of the model. One example, Cook (1979) introduced the Cook's Distance, CD_i. The Cook's distance measure how much the parameter estimates change when a point is remove from the calculation. It is given by

$$CD_i = \frac{\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)^T (X^T X)\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)}{k\hat{\sigma}^2} \tag{2.24}$$

where $\hat{\beta}^{(-i)}$ is the estimated parameter of β when the ith observation is deleted, and k are independent variables in the model.

DFFITS is usually used in regression model to show the influential point. A low leverage point can be detected from a small value of DIFFITS. This measure is defined as:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}} \text{ for } i = 1,2,\dots,n \tag{2.25}$$

where $\hat{y}_i^{(-i)}$ are fitted responds, $\hat{\sigma}_{(i)}$ are the estimated standard error when the ith observation is deleted and h_{ii} is the leverage.

Meanwhile, DFBETA is used to measure the change in each parameter estimate. It is given by

$$DFBETAS_j = \frac{b_j - b_{(i)j}}{s_{(i)}\sqrt{(X'X)_{jj}}} \tag{2.26}$$

where $(X'X)_{jj}$ is the $(j,j)^{th}$ elements of $(X'X)^{-1}$. The DFBETAS is calculated by deleting the ith observation and the large value of DFBETAS showed that the observations are influential.

On the other hand, another diagnostic outlier that is usually used in regression modelling is the *COVRATIO* which has been introduced by Belsley et al. (1980).

COVRATIO statistic identify the change in the determinant of the covariance matrix of the estimates by deleting the i^{th} observation and is defined by

$$COVRATIO_{(-i)} = \frac{|COV|}{\left|COV_{(-i)}\right|} \tag{2.27}$$

where $|COV|$ is the determinant of covariance matrix of full data set and $\left|COV_{(-i)}\right|$ is that of the reduced data set by excluding the i^{th} row. If the ratio is close to the one, this means that there is no significant difference between the covariance matrices. This means, the observation is consistent with the other observations (Ghapor et al., 2014).

Ibrahim et al. (2013) and Rambli et al. (2016) used *COVRATIO* procedure in identifying outliers in circular regression model namely JS circular regression model and DM circular regression model respectively. Moreover, several authors have been used *COVRATIO* statistic in errors-in-variables model focusing on circular data. For example, Hussin et al. (2010) proposed *COVRATIO* statistic in circular functional relationship model. Later, Hussin and Abuzaid (2012) used *COVRATIO* statistic in circular functional relationship model by transforming the circular data using complex form and Mamun et al. (2019) in unreplicated linear structural relationship model. Recently, Mokhtar et al. (2022) and Mokhtar et al. (2019) have been applied *COVRATIO* statistic for circular linear functional relationship model for data in United Kingdom and Malaysia respectively. However, in linear functional relationship model, methods of identifying outliers are somewhat limited. Ghapor et al. (2014) has been used the *COVRATIO* statistic in detecting the outliers for unreplicated linear functional relationship model. Noting the wide applicability of *COVRATIO* statistic, it will be considered in replicated linear

functional relationship model in detecting outliers. Furthermore, its popularity may be due to its simplicity. A detailed discussion is given in Chapter 6.

2.4.2 Robust Statistics

Robust statistics is needed to address the outlier problems in data set. In the traditional procedures, one has to follow the assumptions in order to perform well with no departures. However, with the presence of outliers, this is not possible as it results in spuria estimates. On the other hand, the robust procedures can work more efficiently compared to the traditional procedures when there is a small departure from them. The term robustness refers to the lack of sensitivity relating to small deviations from the assumptions. This means the use of robust statistics is to manage with outliers by keeping the effects of their presence minimal.

There is a large amount of literature available for robust statistics. However, much of the current literature on robust statistics pays particular attention on linear regression (Rousseeuw & Leroy, 1987; Maronna et al., 2006; Hampel et al., 2011). The main application of robust techniques in a regression problem is to develop estimators that are not strongly affected by the presence of the outliers. There a few authors have been discussed the robust statistics in errors-in-variable model; for example, Zamar (1989) proposed orthogonal *M*-estimators for the no-equation error model and found that the *M*-estimators is robust when the measurement errors are elliptically distributed. In structural errors-in-variable model; Fekri and Ruiz-Gazen (2004) proposed robust weighted orthogonal regression and K. M. Jung (2007) proposed an orthogonal least trimmed squares (OLTS) estimator and showed that the proposed estimates are robust and

efficient. Recently, Mamun et al., (2020) proposed a modified maximum likelihood estimators in linear structural relationship model. In this study, the works of Mamun et al., (2020) will be considered in the unreplicated linear functional relationship model to find the robust estimator in the presence of outliers. A detailed explanation is given in Chapter 3.

As mentioned earlier, there is a relatively small body of literature that is concerned with robust statistics in linear functional relationship model. Abdullah (1989) proposed some median-based estimators in linear functional model and showed Theil-type estimators and the modified *L1-norm* estimators are found to robust. In 2005, Al-Nasser and Ebrahem established a nonparametric method using median to estimate slope parameter and showed that his proposed method is more robust than modified least squares, repeated median, geometric mean and Housner and Brennan's (1948) estimator. Also, Ghapor et al. (2015) added another step in Al-Nasser and Ebrahem (2005) method to estimate the slope parameters and showed that the proposed method is more robust than that of Al-Nasser and Ebrahem method. Thus, in this study, the robust technique using the nonparametric method will be explored in replicated linear functional relationship model in estimating the slope parameter and will be discussed in detail in Chapter 5.

2.5 Cluster Analysis

Clustering is a mathematical technique to create or classify groups of similar observations into subsets or cluster based upon a specific algorithm. The algorithms' objective is to make the observations under the same group or cluster are similar to each

other while the observations in different group or cluster are dissimilar from entities in another. Clustering analysis dates back in the 1960s in the field of biology by endorsing the idea of biological taxonomy and also by considering different aspects of cluster analysis techniques (Sokal, 1963). Since then, many authors have been used clustering analysis by different name such as numerical taxonomy in biology, Q analysis in psychology, unsupervised pattern recognition in artificial intelligence and also segmentation in marketing (Everitt et al. 2011). The discussion on clustering analysis namely the meaning of clustering, the clustering methods and the used of software can be found in Blashfield and Aldenderfer (1978). They also commented on future development in cluster analysis by mentioning that consolidation between diverse fields of study is possible, although it is unlikely to occur for a variety of reasons. More on the fundamental concepts and techniques can be found in Jain, Murty, and Flynn (1999).

Johnson and Wichern (2015) mentioned that the objective of clustering is to discover the natural groupings of the variables or the observations. In other words, there is no assumption about the number of groups and the groups are established based on similarities or dissimilarities of the observations. It is necessary to cluster the data when there is no other information or labelled available on the data (Warren Liao, 2005). Clustering is a subjective process in grouping observations based on their similarity measure (Jain et al., 1999). Cluster analysis has been applied to understand data from diverse fields such as in biology, astronomy, social sciences, marketing, geography and many more (Blashfield & Albenderfer, 1978; Everitt et al., 2011; Hartigan, 1975). As mentioned by Kaufman and Rousseeuw (1990), clustering have been used in different fields of studies such as in marketing by identifying market segments, in geography by grouping the regions, in history by grouping the archeology findings to name a few. Jain et al. (1999) have discussed cluster analysis in the form of the pattern representation, similarity computation, grouping process and cluster representation. The application of

cluster analysis can be seen in many disciplines such as in astronomy, biomedical, data science and many other disciplines (Hartigan, 1975). For example, in biomedical study, clustering can be used in microarray gene expression, MRI data analysis, genomic sequence analysis and many others (Xu & Wunsch, 2010). Warren Liao (2005) had listed all datasets used in his survey with several applications in business, engineering, medicine, entertainment and others.

Clustering methods can be divided into two categories namely the hierarchical and non-hierarchical methods. Agglomerative and divisive are the two different classes under hierarchical methods while partitioning, density-based and grid-based are different types under non-hierarchical methods. Each of these methods offers different perspectives on the discovering natural groups in data and the results obtained can be very different when different methods are applied on the same data (Everitt et al., 2011).

Another application can be seen in the formed of data reduction, data summarization, prediction based on groups, finding the nearest neighbours and also in outlier detection (Y. Jung et al., 2003; Van Aelst et al., 2006; Partovi Nia & Davison, 2015; Mokhtar et al., 2017). As mentioned earlier, several authors have been used clustering analysis in their model. Van Aelst et al. (2006) proposed a method called linear grouping algorithm using orthogonal regression to determine the number of groups in data in which can be used in linear functional relationship model. Y. Jung et al. (2003) proposed a method to measure clustering optimality quantitatively with a purpose to use it to determine an optimal number of clusters in various clustering algorithms. Mokhtar et al. (2017) used clustering in identifying the multiple outliers in functional relationship model for circular data. Partovi Nia and Davison (2015) proposed a method based on a mixture model using hierarchical clustering to cluster the replicated and unreplicated data.

2.5.1 Similarity Measure

As defined by Sebert et al., (1998), cluster analysis begins by taking a set of *n* observations on *p* variables where a similarity measure between observations is obtained based on their inter-observation similarities. There are two primary decisions before applying the clustering algorithms namely the measure of similarity to use and also which the clustering algorithm to use. There are four types of similarity measure to group the variables or observations into their own groups namely correlation coefficient, distances measures, association coefficients and probabilistic similarity coefficients (Blashfield & Aldenderfer, 1978).

The most commonly used to compute the distances measures is based on metric function. The purpose of the metric function is to give some ways to measure the observations and their distance in order to decide which elements belong to a group. There are several ways to compute the distance between any pair of points or observations such as the Euclidean distance, the Manhattan distance, the Minkowski distance, Mahalanobis distance and many other distances (Di & Satari, 2017; Murtagh & Contreras, 2012). The Euclidean distance is the most popular common distances measures and also referred straight-line distance which can be defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \tag{2.28}$$

where d_{ij} is the distance between i and j, and x_{ik} is the value of the kth variable for the ith observation and x_{jk} is the value of the kth variable for the jth observation where $i = 1,2, \dots, d$ and $j = 1,2, \dots, d$.

Another type of measurement distance is the Manhattan distance, or known as a city-block metric. It represents the distance between points in a city road grid, which is defined as

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|^{r} \tag{2.29}$$

where d_{ij} is the distance between i and j, and x_{ik} is the value of the kth variable for the ith observation and x_{jk} is the value of the kth variable for the jth observation where $i = 1,2, \dots, d$ and $j = 1,2, \dots, d$.

The Minkowski distance is a special class of metric distance because it is a generalisation of the Euclidean and Manhattan distance with different values of r. The Minkowski distance can be defined as

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{r} \right)^{\frac{1}{r}} \tag{2.30}$$

where d_{ij} is the distance between i and j, and x_{ik} is the value of the kth variable for the ith observation and x_{jk} is the value of the kth variable for the jth observation where $i =$

$1,2,\dots,d$ and $j = 1,2,\dots,d$. If $r = 1$, this distance becomes the Manhattan distance and if $r = 2$, the distance becomes the Euclidean distance.

Alternatively, the other distance is called Mahalanobis distance or known as generalized distance that accounts for the correlations among variables in a way that weights each variable equally which is defined as

$$d_{ij} = (X_i - X_j)\sum^{-1}(X_i - X_j) \tag{2.31}$$

where $\sum^{-1}$ is the pooled within-groups variance-covariance matrix, and X_i and X_j are vectors of the values of the variables for observation *i* and *j*.

In this study, the Euclidean distance will be used as the similarity measure as defined in (2.28) because not only it is frequently used but simple to apply and widely recognized when grouping multivariate observations (Everitt et al., 2011; Johnson & Wichern, 2015). The characteristic in Euclidean distance is that the relatively small distance should separate similar observations, whereas a relatively larger distance should separate dissimilar observations.

2.5.2 Agglomerative Hierarchical Clustering

Hierarchical clustering is the most popular used algorithm as it is simple and easy to use (Dasgupta & Long, 2005). In hierarchical clustering, there is no prior specification of the number of clusters as it does not require to pre-specify the number of clusters to be

generated. The hierarchical cluster works on similarity matrix to construct a tree representing specified relationship between observations by dividing the observations into two groups, that is the agglomerative and the divisive.

The agglomerative methods build tree from branches to root, while the divisive methods begin at the root and work toward the branches. The agglomerative hierarchical method starts with a series of successive merging between individual observations as clusters. First, the most similar objects are grouped initially based on their similarity measure. As the similarity decreases, in the end, all the subgroups are fused into a single cluster. These clusters are permanently merged together or nested. As for the divisive hierarchical methods, the initial group consist of all the objects. Then, the group are divided into two subgroups in a such the observations in one group are distant from the observations in the other. The process of division will continue until there are many subgroups of the observations. The results from both the agglomerative and divisive hierarchical clustering may be displayed in the form of a dendogram or usually define as the tree diagram. Figure 2.1 shows an example of the dendogram that illustrates the root and branches in a hierarchical clustering.

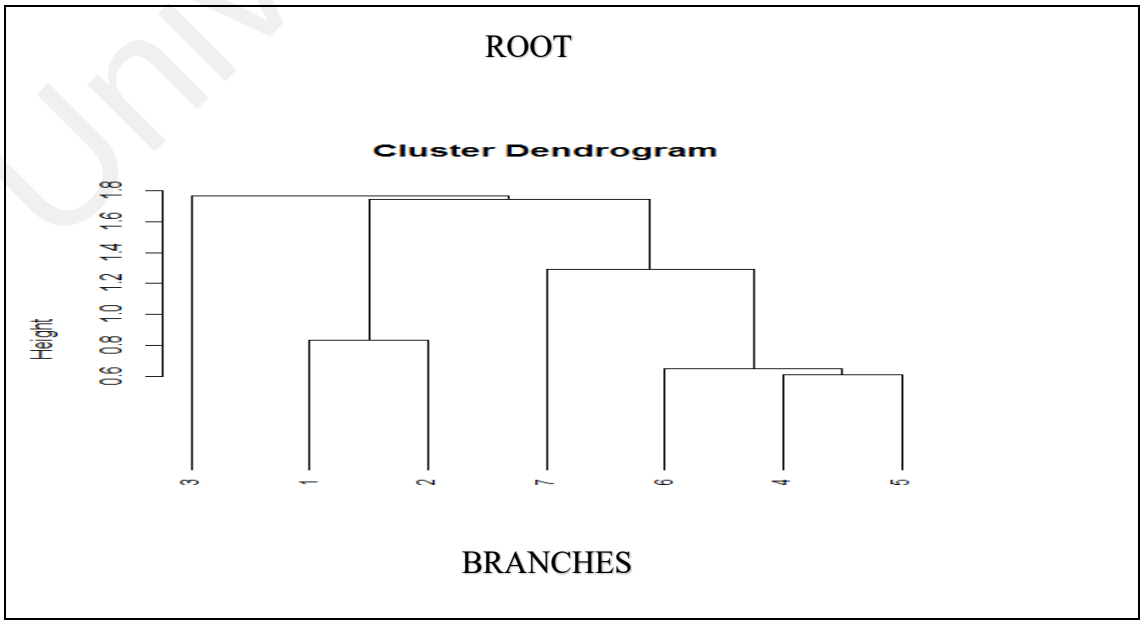

Figure 2.1 Illustration of branches and root in hierarchical clustering

From Figure 2.1, each connected observations forms a cluster by cutting the dendogram at the desired level or by setting the termination condition. The termination conditions and stopping rule had been reviewed by Milligan and Cooper (1985) and Y. Jung et al. (2003). Murtagh and Contreras (2012) reviewed the hierarchical clustering and gave some recommendations on algorithmic aspects and computational properties.

According to Kaufman and Rousseeuw (1990), there are several major clustering techniques in agglomerative hierarchical clustering and are summarized as follows:

1. Single linkage method defines as the similarity between clusters as the shortest distance from any object in one cluster to any object in the second cluster.
2. Complete linkage method defines as the similarity between clusters as the largest distance from any object in one cluster to any object in the second cluster.
3. Average linkage method defines as the similarity between clusters as the average distance from any object in one cluster to any object in the second cluster.
4. Centroid method used the similarity between two clusters is the distance between the cluster centroids.
5. Ward's method used the similarity between two clusters is the sum of squares within the clusters summed over all variables.

In this study, the agglomerative hierarchical clustering namely the single linkage, the average linkage and the complete linkage will be considered as the method of grouping the data from unreplicated data to the replicated data. The number of group or cluster will be used as the termination condition. A detail discussion on this topic is given in Chapter 7.

2.6 Table of Summary

Table 2.1 and Table 2.2 showed some of the main authors and their findings that have been frequently used in this research including the parameter estimation in linear functional relationship model and outlier detection using *COVRATIO* statistic.

Table 2.1 Literature Review on Linear Functional Relationship Model

Author (Year)	Title	Main Findings
Lindley (1947)	Regression Lines and the Linear Functional Relationship	Proposed the ratio of error variances need to be known in order to overcome the inconsistencies in linear functional relationship model.
Dorff and Gurland (1961)	Estimation of the Parameters of a Linear Functional Relation	Compared various consistent slope estimators for both unreplicated and replicated linear functional relationship model.
Barnett (1970)	Fitting Straight Lines – The Linear Functional Relationship with Replicated Observations	Derived maximum likelihood estimation method in replicated linear functional relationship model but the model can be improved by choosing another

		suitable initial value and iteration process.
Dolby et al. (1987)	On Fitting Bivariate Functional Relationships to Unpaired and Unequally Replicated Data	Proposed unpaired and unequally replicated linear functional relationship model based on error variances.
Hussin et al. (2005)	Pseudo-replicates in the Linear Circular Functional Relationship Model	Derived the parameter estimation using maximum likelihood estimation method for replicated circular functional relationship model.
Mokhtar et al. (2017)	On Parameter Estimation of a Replicated Linear Functional Relationship Model for Circular Variables.	Proposed estimation of the rotation parameter in replicated linear functional relationship model on circular variables.

Table 2.2 Literature Review on Outlier Detection using *COVRATIO* statistic

Author (Year)	Title	Findings
Belsley et al. (1980)	Identifying Influential Data and Sources of Collinearity	Introduced *COVRATIO* statistic in regression modelling.
Hussin et al. (2010)	Asymptotic Covariance and Detection of Influential Observations in Linear Functional Relationship Model for Circular Data with Application to the Measurements of Wind Directions.	Proposed outlier detection method using *COVRATIO* statistic in a linear functional relationship model for circular data.
Abuzaid et al. (2011)	*COVRATIO* Statistic for Simple Circular Regression Model	Proposed outlier detection method using *COVRATIO* statistic in simple circular regression model.
Ibrahim et al. (2013)	Outlier Detection in a Circular Regression Model using *COVRATIO* Statistic	Proposed outlier detection method using *COVRATIO* statistic in JS circular regression model.
Ghapor et al. (2014)	On Detecting Outlier in Simple Linear Functional Relationship Model using *COVRATIO* Statistic	Proposed outlier detection using *COVRATIO* statistic in unreplicated linear functional relationship model.
Rambli et al. (2016)	Outlier Detection in a Circular Regression Model	Proposed outlier detection using *COVRATIO* statistic in DM circular regression model.

Mamun et al. (2019)	Identification of Influential Observation in Linear Structural Relationship Model with Known Slope	Proposed outlier detection method using *COVRATIO* statistic in unreplicated linear structural relationship model.
Mokhtar et al. (2019)	An Outlier Detection Method for Circular Linear Functional Relationship Model using *COVRATIO* Statistic	Proposed outlier detection method using *COVRATIO* statistic in a linear functional relationship model for circular data with equal error concentration parameters.

CHAPTER 3: MODIFIED MAXIMUM LIKELIHOOD ESTIMATION FOR THE SLOPE PARAMETERS OF UNREPLICATED LINEAR FUNCTIONAL RELATIONSHIP MODEL

3.1 Introduction

In this chapter, a modified maximum likelihood estimation method for the unreplicated linear functional relationship model (LFRM) is proposed. In this case, the ratio of error variances is assumed known and equal to one ($\lambda = \frac{\tau^2}{\sigma^2} = 1$). The motivation of this proposed method is to introduce a robust method in the estimation of the slope parameter in the presence of outliers. This chapter aims to address the first objective of the study namely to develop a modified maximum likelihood estimation for the slope parameter in unreplicated linear functional relationship model particularly in the presence of outliers. The organization of the chapter is as follows. In Section 3.2, the parameter estimation using the traditional method namely the maximum likelihood estimation method is discussed. The modified maximum likelihood estimation method is proposed and described in detail in the Section 3.3. In order to measure the robustness of the maximum likelihood method and modified maximum likelihood method, simulation studies are carried out in Section 3.4. The results and discussion are presented in Section 3.5. In Section 3.6, practical examples using real datasets are also presented. Summary and conclusions are given in Section 3.7.

3.2 Maximum Likelihood Estimation Method

As mentioned earlier, this section describes the maximum likelihood method to estimate parameters in linear functional relationship model in the case of known error variance ratio, namely $\tau^2 = \lambda\sigma^2$ and the value of the ratio, λ is equal to one (Lindley, 1947).

Suppose X_i and Y_i are linearly related but observed with error with the equation

$$Y_i = \alpha + \beta X_i \tag{3.1}$$

where α is the intercept parameter and β is the slope parameter. Assume that for each x_i and y_i are subjected to errors δ_i and ε_i instead of X_i and Y_i respectively and $i = 1,2, \dots, n$. Then it can be modelled as,

$$x_i = X_i + \delta_i \text{ and } y_i = Y_i + \varepsilon_i \tag{3.2}$$

The error terms δ_i and ε_i are assumed to be mutually independent and normally distributed random variables, that is, $\delta_i \sim N(0, \sigma^2)$ and $\varepsilon_i \sim N(0, \tau^2)$. This model is also known as unreplicated linear functional relationship model since there is only a single observation for each level i with the assumption the ratio of error variances is equal to one, that is $\lambda = 1$.

The log-likelihood function for the parameters of unreplicated linear functional relationship model can be written as

$$\log L\left(\alpha, \beta, \sigma^2, X_1, \ldots, X_n; \lambda, x_1, \ldots, x_{n,}, y, \ldots, y_{n,}\right) = \tag{3.3}$$

$$-n\ log(2\pi) - \frac{n}{2}\log\lambda - n\log(\sigma^2) - \frac{\sum(x_i - X_i)^2}{2\sigma^2} - \frac{\sum(y_i - \alpha - \beta X_i)^2}{2\lambda\sigma^2}$$

There are (n +3) parameters to be estimated which are the intercept, α, the slope, β, the error variance, σ^2, and the incidental parameters, $X_1, \cdots, X_n$, as in (3.3). However, in this study, our primary interest is on the slope parameter, β, and thus is the scope of the study. These parameters in unreplicated linear functional relationship model may be obtained by differentiating the log likelihood function with respect to α, β, σ^2 and X_i respectively and equating to zero. Thus, the estimates $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$ and $\hat{X}_i$ can be obtained given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{3.4}$$

$$\hat{\beta} = \frac{\left(S_y^2 - \lambda S_x^2\right) + \sqrt{\left(S_y^2 - \lambda S_x^2\right)^2 + 4\lambda S_{xy}^2}}{2S_{xy}}, \tag{3.5}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left\{\sum_{i=1}^{n}(x_i - X_i)^2 + \frac{1}{\lambda}\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}X_i\right)^2\right\}, \text{ and} \tag{3.6}$$

$$\hat{X}_i = \frac{\lambda x_i + \hat{\beta}\ (y_i - \hat{\alpha})}{\lambda + \hat{\beta}^2} \tag{3.7}$$

where $S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$, $S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$ and $S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$ respectively.

However, the presence of outliers will affect the parameter estimation of unreplicated linear functional relationship model especially the slope parameter (Al-Nasser & Ebrahem, 2005; Ghapor et al., 2015). From (3.5**)**, the equation of the slope parameter are depending on the sample mean, the sample variance and the sample covariance which are sensitive to the outliers even a single outlier (Maronna et al., 2006). Thus, a robust method is needed to overcome this problem.

3.3 Modified Maximum Likelihood Estimation Method

As mentioned earlier, it is important to modify parameter estimation using maximum likelihood estimation method when outliers are present in dataset. In this section, a modification to the estimation of the slope parameter of the unreplicated linear functional relationship model is proposed. For this model, the ratio of error variances is known and is equal to one, $\lambda = \frac{\tau^2}{\sigma^2} = 1$, to overcome the unidentifiability problem in linear functional relationship model.

It is well established that standard statistics such as sample mean, sample variance and sample covariance in the original parameter estimation are not robust to outliers, thus the modification is deemed necessary. Several authors modified the classical estimators by

modifying the traditional method with robust estimators (Koláček, 2008; Liu, 2012). The proposed method namely the modified maximum likelihood estimation method should be able to produce a reasonable estimate even in the presence of outliers. Although the proposed method has been used in linear structural relationship model (LSRM) in estimating the intercept and the slope parameters (Mamun et al., 2020), the differences between both linear structural relationship model (LSRM) and linear functional relationship model (LFRM) is the incidental parameter where in linear structural relationship model, X_i is random but in linear functional relationship model, the X_i is fixed. Thus, by getting the robust estimator of the slope, β, the other parameters also can be estimated such as the intercept, α, the error variance, $\hat{\sigma}^2$, and the incidental parameter, X_i in the presence of outliers although our primary interest is the slope parameter, β, of unreplicated linear functional relationship model. The aim of robust methods is to ensure high stability of statistical inference under the deviations from the assumed distribution model (Shevlyakov & Smirnov, 2011). In short, in order to overcome the outlier problem, a modified maximum likelihood estimation (MMLE) method by replacing the usual estimators with robust estimator is proposed.

The following are the estimates. For measure of central tendency, it is well established that median is more robust measure than mean. Thus, in this study, it is proposed that the sample mean is replaced by the sample median instead and denoted as follows:

$$\bar{x}_{Rob} = median\,(x), \qquad \bar{y}_{Rob} = median\,(y) \tag{3.8}$$

and the sample variances are replaced by a robust estimator,Q_n .

$$\{Q_n(x)\}^2 = S_x^2(Rob), \qquad \{Q_n(y)\}^2 = S_y^2(Rob) \tag{3.9}$$

where,

$$Q_n(x) = 1.0483\{|x_i - x_j; i < j|\}_{(k)}, Q_n(y) = 1.0483\{|y_i - y_j; i < j|\}_{(k)}$$

and $k = \binom{h}{2} \approx \binom{n}{2}/4$ where $h = (n/2)+1$ is a roughly half the number of observations.

The robust estimator, Q_n is suitable for asymmetric distribution and with a 50% breakdown point (Rousseeuw and Croux, 1993).

The sample covariance S_{xy} is denoted by $S_{xy}(Rob)$, in which,

$$S_{xy}(Rob) = r_{Qn} \times S_x(Rob) \times S_y(Rob) \tag{3.10}$$

where r_{Qn} is the robust correlation coefficient proposed by Shevlyakov and Smirnov (2011) and defined as

$$r_{Qn} = \frac{\{Q_n(u)\}^2 - \{Q_n(v)\}^2}{\{Q_n(u)\}^2 + \{Q_n(v)\}^2} \tag{3.11}$$

where *u* and *v* are the robust principle defined as

$$u = \frac{x-median(x)}{\sqrt{2}\,Q_n(x)} + \frac{y-median(y)}{\sqrt{2}\,Q_n(y)} \text{ and } v = \frac{x-median(x)}{\sqrt{2}\,Q_n(x)} - \frac{y-median(y)}{\sqrt{2}\,Q_n(y)}$$

In order to obtain the modified maximum likelihood estimation for the slope of unreplicated linear functional relationship model, the original estimators in maximum likelihood estimation is replaced by the robust estimators as mentioned above. Thus, the new slope, $\hat{\beta}_{MMLE}$, is as follows:

$$\hat{\beta}_{MMLE} = \frac{\left(S_y^2(Rob) - \lambda S_x^2(Rob)\right) + \sqrt{\left(S_y^2(Rob) - \lambda S_x^2(Rob)\right)^2 + 4\lambda S_{xy}^2(Rob)}}{2S_{xy}(Rob)} \tag{3.12}$$

As a reference, Mamun et al. (2020) proposed a modified maximum likelihood estimation method but for a LSRM instead.

3.4 Simulation Studies

Simulation studies was carried out using the R software in order to evaluate the performance of the proposed method namely the modified maximum likelihood estimation method, together with the existing method, namely the maximum likelihood estimation method, for unreplicated linear functional relationship model in the presence of the outliers. The study begins by simulating the data from unreplicated linear functional relationship model given by:

$$x_i = X_i + \delta_i \quad , \quad y_i = Y_i + \varepsilon_i \tag{3.13}$$

$$\text{and } Y_i = 1 + X_i \tag{3.14}$$

where $X_i = 10\frac{i}{n}$, $i = 1, \cdots, n$ and X_i is a fixed constant. The error terms δ_i and ε_i , both are taken from Normal distribution with mean 0 and variances 0.1; $\delta_i \sim N(0,0.1)$ and $\varepsilon_i \sim N(0,0.1)$.

As mentioned by Hampel et al. (2011), the data set could contain about 1% to 10% outliers although in some literature, authors consider outliers up to 50% to obtain robust parameters. However, this study considered data with no outlier, a single outlier, 10%, 20% and 30% outliers respectively. The rationale is, if the data set consists of more than 50% outliers, then it will be hard to differentiate between the actual observations and the outliers. According to Al-Nasser and Ebrahem (2005), to contaminate data points, for example at point *c* for variable *y* can be done by using the relationship,

$$y_c = 1 + X_c + \varepsilon_c \text{ where } \varepsilon_c \sim N(\mu, \sigma^2). \tag{3.15}$$

In this study, the error term ε_c is taken from Normal distribution with mean 0 and variance 25; $\varepsilon_c \sim N(0,25)$ and has replaced the original observations, y with contaminated observations, y_c.

In each trials, sample size of 20,50,80 and 100 are generated using relationship in equation (3.13) and (3.14). In order to investigate the robustness of proposed method, the error terms δ_i and ε_i are generated from non-normal distribution namely beta distribution. The probability density function of the beta distribution as follows:

$$f(x) = \frac{1}{B(a,b)} x^{(a-1)}(1-x)^{(b-1)}, \qquad 0 \leq x \leq 1 \tag{3.16}$$

where a and b are two positive shape parameters, and $B(a,b)$ can be defined as the beta function. In this study, three different beta distributions are considered namely beta distribution with parameters (2,9) for right-skewed case, beta distribution with parameters (9,2) for left-skewed case and beta distribution with parameters (3,3) for non-normal symmetric case. The graph of probability density function for the parameters mentioned above can be shown in the Figure 3.1 below.

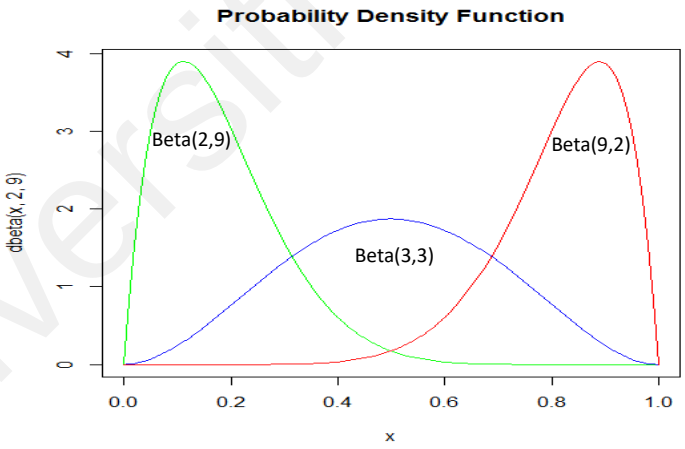

Figure 3.1 The Probability Density Function for Beta Distribution

The performance used in the simulation study is based on estimated bias (EB) and mean square error (MSE) using 10000 trials. The estimated bias and the mean square error are defined by

$$\text{Estimated Bias, EB} = |\hat{w} - w| \text{ and} \tag{3.17}$$

$$\text{Mean Square Error, MSE} = \frac{1}{s}\sum\left(\hat{w}_j - w\right)^2 \tag{3.18}$$

where w be a generic term for the parameters and s is a number of simulation.

3.5 Results and Discussion

Results of the performance based on estimated bias of the proposed method namely modified maximum likelihood estimation method and the traditional method, maximum likelihood estimation method in estimating the slope parameter are shown in Table 3.1-Table 3.4 respectively. From the simulation results in Table 3.1, when the error terms are from the Normal distribution i.e. for the normal case, there is no much difference between the maximum likelihood estimation method and the modified maximum likelihood estimation method in estimating the slope parameter when the data free from outlier. The estimated bias (EB) of the proposed method, modified maximum likelihood estimation method and the traditional method, maximum likelihood estimation method is somewhat similar to each other. However, when a single outlier or multiple outliers present, the estimated bias for modified maximum likelihood estimation method is less than the maximum likelihood estimation method in estimating the slope parameter as the sample size increases from 20 to 100. Similar results can be observed for non-normal distribution which in this case is from beta distribution namely the right skewed case, the left skewed case and the non-normal symmetric case with parameter (2,9), (9,2) and (3,3) as shown in Table 3.2, Table 3.3 and Table 3.4 respectively. This shows that the modified maximum likelihood estimation method is a superior method in estimating the slope parameter

compared to the maximum likelihood estimation method in the presence of outliers even though there is a single outlier presence in the dataset.

Table 3.1 Estimated Bias of the Slope: Normal Case: Normal (0,0.1)

Outliers	Method	$n = 20$	$n = 50$	$n = 80$	$n = 100$
No outlier	MLE	1.9350E-04	9.7890E-06	2.9262E-05	2.3213E-05
	MMLE	1.5991E-03	1.8946E-04	5.6520E-05	7.5797E-06
Single	MLE	6.6469E+00	8.0430E-01	3.9753E-01	2.9533E-01
	MMLE	6.5918E-02	2.2669E-02	1.4304E-02	1.1466E-02
10%	MLE	1.2604E+01	1.2657E+01	1.2658E+01	1.2662E+01
	MMLE	1.0215E-01	9.0805E-02	8.7843E-02	8.6883E-02
20%	MLE	6.3234E+00	6.3293E+00	6.3301E+00	6.3306E+00
	MMLE	1.3382E-01	1.1291E-01	1.0572E-01	1.0373E-01
30%	MLE	5.6076E+00	5.6116E+00	5.6118E+00	5.6120E+00
	MMLE	1.3367E-01	1.1297E-01	1.0577E-01	1.0372E-01

Table 3.2 Estimated Bias of the Slope: Right Skewed Case: Beta (2,9)

Outliers	Method	$n = 20$	$n = 50$	$n = 80$	$n = 100$
No outlier	MLE	2.8112E-04	3.9492E-05	9.8312E-05	4.9143E-05
	MMLE	8.8786E-05	9.5760E-05	2.7095E-04	3.2893E-05
Single	MLE	6.6657E+00	8.0454E-01	3.9744E-01	2.9536E-01
	MMLE	6.5428E-02	2.2729E-02	1.4505E-02	1.1401E-02
10%	MLE	1.2610E+01	1.2658E+01	1.2668E+01	1.2664E+01
	MMLE	1.0296E-01	9.0423E-02	8.8167E-02	8.6928E-02
20%	MLE	6.3216E+00	6.3287E+00	6.3301E+00	6.3300E+00
	MMLE	1.3496E-01	1.1297E-01	1.0626E-01	1.0379E-01
30%	MLE	5.6072E+00	5.6113E+00	5.6118E+00	5.6121E+00
	MMLE	1.3478E-01	1.1307E-01	1.0647E-01	1.0398E-01

Table 3.3 Estimated Bias of the Slope: Left Skewed Case: Beta (9,2)

Outliers	Method	$n = 20$	$n = 50$	$n = 80$	$n = 100$
No outlier	MLE	2.0058E-04	6.4640E-05	9.5621E-05	6.4618E-05
	MMLE	7.3848E-04	1.2247E-04	3.1907E-04	1.2004E-05
Single	MLE	6.6626E+00	8.0419E-01	3.9744E-01	2.9536E-01
	MMLE	6.6007E-02	2.2718E-02	1.3905E-02	1.1440E-02
10%	MLE	1.2609E+01	1.2663E+01	1.2658E+01	1.2663E+01
	MMLE	1.0346E-01	9.0591E-02	8.7590E-02	8.6925E-02
20%	MLE	6.3220E+00	6.3306E+00	6.3304E+00	6.3313E+00
	MMLE	1.3562E-01	1.1315E-01	1.0602E-01	1.0396E-01
30%	MLE	5.6076E+00	5.6122E+00	5.6122E+00	5.6124E+00
	MMLE	1.3574E-01	1.1339E-01	1.0591E-01	1.0383E-01

Table 3.4 Estimated Bias of the Slope: Non-normal Symmetric Case: Beta(3,3)

Outliers	Method	$n = 20$	$n = 50$	$n = 80$	$n = 100$
No outlier	MLE	2.9325E-04	3.1923E-04	3.2210E-05	8.1810E-05
	MMLE	4.7079E-04	3.7636E-04	1.8625E-04	2.0192E-04
Single	MLE	6.7192E+00	8.0589E-01	3.9743E-01	2.9541E-01
	MMLE	5.6750E-02	2.3240E-02	1.4433E-02	1.1609E-02
10%	MLE	1.2694E+01	1.2702E+01	1.2683E+01	1.2688E+01
	MMLE	1.0164E-01	9.1176E-02	8.8394E-02	8.7403E-02
20%	MLE	6.3266E+00	6.3312E+00	6.3315E+00	6.3326E+00
	MMLE	1.4725E-01	1.1567E-01	1.0918E-01	1.0699E-01
30%	MLE	5.6094E+00	5.6128E+00	5.6128E+00	5.6127E+00
	MMLE	1.4865E-01	1.1648E-01	1.0933E-01	1.0701E-01

As mentioned earlier, another performance measure of mean square error is also used. Looking at Table 3.5 when the error terms are taken from normal distribution, the mean square error (MSE) of traditional method and the proposed method give similar result when no outlier exists in the dataset. However, when contamination level increase, the mean square error of the slope estimator using maximum likelihood estimation method becomes larger. The modified maximum likelihood estimation method, on the other hand, the mean square error is consistently small even when the presence of the outliers increase. This is true for all sample size considered. The values of the mean square error for the proposed modified maximum likelihood estimation method, gave smaller values than the maximum likelihood estimation method as sample size increase from 20 to 100. From Table 3.6, Table 3.7 and Table 3.8, the same can be said for other cases in which

when the error terms are taken from beta distribution namely the right-skewed case, beta (2,9), the left-skewed case, beta (9,2) and the non-normal symmetric case, beta (3,3) respectively.

Table 3.5 Mean Square Error of the Slope: Normal Case: Normal (0,0.1)

Outliers	Method	$n = 20$	$n = 50$	$n = 80$	$n = 100$
No outlier	MLE	1.1826E-04	4.7859E-05	3.0419E-05	2.4422E-05
	MMLE	1.3623E-03	2.0496E-04	1.0386E-04	7.4214E-05
Single	MLE	4.4369E+01	6.4742E-01	1.5815E-01	8.7298E-02
	MMLE	5.6066E-03	7.2873E-04	3.1114E-04	2.0728E-04
10%	MLE	1.5929E+02	1.6038E+02	1.6035E+02	1.6043E+02
	MMLE	1.1688E-02	8.4999E-03	7.8489E-03	7.6427E-03
20%	MLE	3.9998E+01	4.0067E+01	4.0076E+01	4.0083E+01
	MMLE	1.9302E-02	1.3047E-02	1.1331E-02	1.0874E-02
30%	MLE	3.1452E+01	3.1495E+01	3.1496E+01	3.1498E+01
	MMLE	1.9355E-02	1.3104E-02	1.1367E-02	1.0893E-02

Table 3.6 Mean Square Error of the Slope: Right Skewed Case: Beta (2,9)

Outliers	Method	$n = 20$	$n = 50$	$n = 80$	$n = 100$
No outlier	MLE	1.5092E-04	6.0035E-05	3.6931E-05	2.9451E-05
	MMLE	1.4451E-03	2.3408E-04	1.1825E-04	8.3231E-05
Single	MLE	4.4655E+01	6.4791E-01	1.5811E-01	8.7330E-02
	MMLE	5.6643E-03	7.5974E-04	3.3243E-04	2.1643E-04
10%	MLE	1.5954E+02	1.6046E+02	1.6063E+02	1.6049E+02
	MMLE	1.1955E-02	8.4699E-03	7.9225E-03	7.6647E-03
20%	MLE	3.9977E+01	4.0061E+01	4.0077E+01	4.0075E+01
	MMLE	1.9705E-02	1.3110E-02	1.1468E-02	1.0901E-02
30%	MLE	3.1447E+01	3.1491E+01	3.1496E+01	3.1500E+01
	MMLE	1.9745E-02	1.3185E-02	1.1545E-02	1.0967E-02

Table 3.7 Mean Square Error of the Slope: Left Skewed Case: Beta (9,2)

Outliers	Method	$n = 20$	$n = 50$	$n = 80$	$n = 100$
No outlier	MLE	1.4864E-04	6.0281E-05	3.7803E-05	2.9657E-05
	MMLE	1.4429E-03	2.2892E-04	1.1736E-04	8.3353E-05
Single	MLE	4.4632E+01	6.4736E-01	1.5812E-01	8.7332E-02
	MMLE	5.7167E-03	7.5616E-04	3.1524E-04	2.1725E-04
10%	MLE	1.5950E+02	1.6055E+02	1.6038E+02	1.6046E+02
	MMLE	1.2020E-02	8.4988E-03	7.8198E-03	7.6656E-03
20%	MLE	3.9981E+01	4.0085E+01	4.0081E+01	4.0092E+01
	MMLE	1.9864E-02	1.3149E-02	1.1418E-02	1.0934E-02
30%	MLE	3.1452E+01	3.1502E+01	3.1501E+01	3.1503E+01
	MMLE	1.9989E-02	1.3256E-02	1.1425E-02	1.0934E-02

Table 3.8 Mean Square Error of the Slope: Non-normal Symmetric Case: Beta (3,3)

Outliers	Method	$n = 20$	$n = 50$	$n = 80$	$n = 100$
No outlier	MLE	4.2920E-04	1.6923E-04	1.0738E-04	8.5749E-05
	MMLE	1.8457E-03	4.9125E-04	2.6520E-04	1.9990E-04
Single	MLE	4.5832E+01	6.5115E-01	1.5838E-01	8.7519E-02
	MMLE	5.2842E-03	1.0558E-03	4.8380E-04	3.4056E-04
10%	MLE	1.6268E+02	1.6196E+02	1.6123E+02	1.6128E+02
	MMLE	1.2622E-02	8.9414E-03	8.1435E-03	7.8925E-03
20%	MLE	4.0058E+01	4.0100E+01	4.0099E+01	4.0112E+01
	MMLE	2.4455E-02	1.4130E-02	1.2325E-02	1.1748E-02
30%	MLE	3.1480E+01	3.1511E+01	3.1509E+01	3.1508E+01
	MMLE	2.5161E-02	1.4451E-02	1.2453E-02	1.1822E-02

To summarize, based on the simulation studies, the proposed method is a robust estimator in the presence of outliers, unlike the maximum likelihood estimation method. It can be concluded that reasonable and satisfactory results obtained from simulation studies where smaller estimated bias and mean square error for all of the cases when the error terms are taken from the normal as well as the non-normal distributions. This shows that the proposed method, modified maximum likelihood estimation method is a better method in estimating the slope parameter compared to maximum likelihood estimation method in the presence of outliers.

3.6 Examples

In this section, two datasets are considered to examine the performance of the modified maximum likelihood estimation method. The measurement error is assumed to occur in either or both variables of this experiment from two datasets in order to apply the relationship as in model (3.1).

3.6.1 Fat Mass Measurements Data

The Fat Mass Measurements Data taken from Goran et al. (1996) is considered. The data consists of 96 fat mass measurements taken on the children by using two techniques namely the skinfold thickness (ST), X_i and bioelectrical resistance (BR), Y_i. The measurement error is assumed to occur in both variables. In unreplicated linear functional relationship model, the assumption on ratio of error variances, $\lambda = 1$ is made to estimate the parameters namely the intercept, α, the slope, β and the error variance, σ^2. A detailed description of the data is given in APPENDIX A. The scatter plot of the skinfold thickness (ST) and the bioelectrical resistance (BR) given in Figure 3.2.

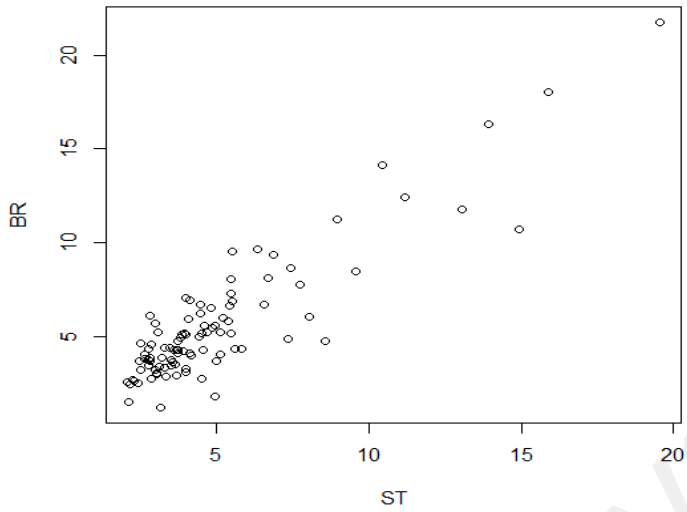

Figure 3.2 The scatter plot of skinfold thickness (ST) and bioelectrical resistance (BR)

As illustrated in Figure 3.2, there exists a linear relationship between these variables and it appears no outliers in the dataset. The data can be modelled by

$$Y_i = \alpha + \beta X_i$$

for $i = 1,2,\cdots,96$ where α is the intercept parameter and β is the slope parameter.

For the unreplicated linear functional relationship model, when the error variance ratio is equal to one, $\lambda = \frac{\tau^2}{\sigma^2} = 1$, the estimated value of the parameters, $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}^2$ using the maximum likelihood method as in (3.4), (3.5) and (3.6) respectively are given in Table 3.9.

Table 3.9 Estimated value of parameters of unreplicated LFRM for Fat Mass Measurements data

Parameters of unreplicated LFRM	Estimated value
$\hat{\alpha}$	0.0787
$\hat{\beta}$	1.0997
$\hat{\sigma}^2$	1.0817

Since the original data do not contain any outlier, the original data is modified following Kim (2000) and Imon and Hadi (2008) by inserting a few outliers to create different cases namely a single outlier, 10%, 20% and 30% outliers. Both the maximum likelihood estimation method and modified maximum likelihood estimation method are applied to estimate the parameters of unreplicated linear functional relationship model on Fat Mass Measurements data and also on modified data. The results are shown in Table 3.10.

Table 3.10 Estimated parameters and standard deviations using two different techniques in Fat Mass Measurements data

Method	MLE		MMLE	
Estimator/ Contamination	Slope	Standard deviation	Slope	Standard deviation
No outlier	1.0997	0.0578	1.2013	0.0635
Single outlier	1.5145	0.2884	1.2358	0.2282
10%	12.2850	13.9307	1.4881	0.7053
20%	81.1245	360.9368	2.0876	1.6520
30%	72.0512	193.7029	2.8885	1.6452

The estimated value of the slope parameter and standard deviation using maximum likelihood estimation method and modified maximum likelihood estimation method is given in Table 3.10. When the data do not have any outliers, the slope parameter for modified maximum likelihood estimation method is slightly difference but comparable to the maximum likelihood estimation method. When the outliers are introduced in the data, the slope parameter in maximum likelihood estimation method breaks down quickly with the increase in the percentage of outliers while the modified maximum likelihood estimation method is not much affected by the existence of outliers.

3.6.2 Frosted Flakes Data

The second example is the frosted flakes data taken from Maindonald and Braun (2010). The data set consisted 100 observations of sugar concentrations (in percentage) for approximately 25 g of cereal samples measured by two techniques, namely the high-performance liquid chromatography (a slow and accurate laboratory method), X_i and a quick method using the infra-analyser 400 (IA400), Y_i. A detailed description of the data can be found in APPENDIX B. The data can be modelled by

$$Y_i = \alpha + \beta X_i$$

for $i = 1,2,\cdots,100$ where α is the intercept parameter and β is the slope parameter. The relationship between the two methods (when the data contains no outlier) used to measure the sugar content in the cereal samples is illustrated in the scatter plot shown below.

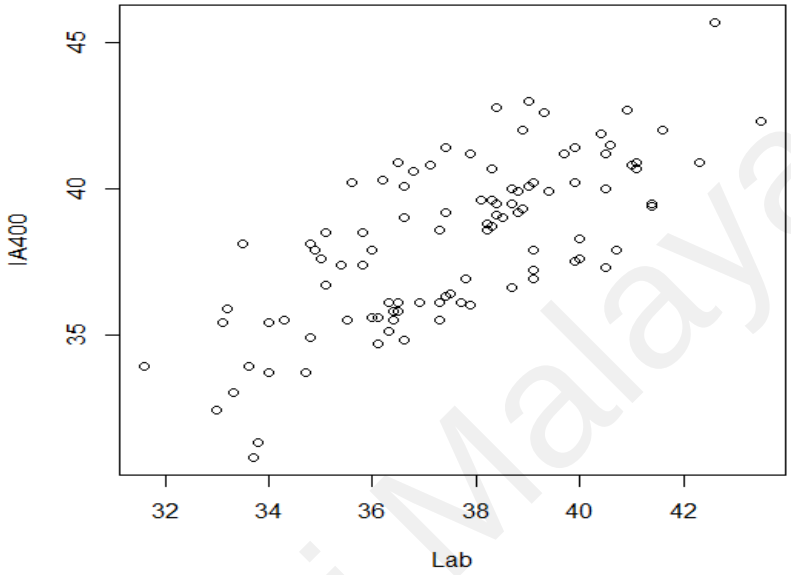

Figure 3.3 The scatter plot for laboratory method (Lab) and a method using the infra-analyser 400 (IA400)

As illustrated in Figure 3.3, there exists a linear relationship between these variables and it appears no outliers in the dataset. For the unreplicated linear functional relationship model, when the error variance ratio is equal to one, $\lambda = 1$, the estimated value of the intercept, the slope and the error variance parameters using the maximum likelihood method is given in Table 3.11.

Table 3.11 Estimated value of parameters of unreplicated LFRM for Frosted Flakes data

Parameters of unreplicated LFRM	Estimated value
$\hat{\alpha}$	0.0787
$\hat{\beta}$	1.0997
$\hat{\sigma}^2$	1.0817

Similar to the previous example, the original data do not contain any outlier. The original data is modified following Kim (2000) and Imon and Hadi (2008) by inserting a few outliers to create different cases namely single outlier, 10%, 20% and 30% outliers. Both the maximum likelihood estimation method and modified maximum likelihood estimation method are applied to estimate the parameters of unreplicated linear functional relationship model on these data and the results are shown in Table 3.12.

Table 3.12 Estimated parameters and standard deviations using two different methods from frosted flakes data

Method	MLE		MMLE	
Estimator/ Contamination	Slope	Standard deviation	Slope	Standard deviation
No outlier	1.1857	0.1602	1.0838	0.1456
Single outlier	1.3603	0.2662	1.0837	0.2075
10%	8.2811	6.0860	1.4204	0.5249
20%	18.1940	14.0503	1.7692	0.4887
30%	79.2880	184.0266	2.5095	1.1153

Table 3.12 shows the advantage of using modified maximum likelihood estimation method. When the data has no outlier, the slope estimate using modified maximum likelihood estimation method is quite similar to the maximum likelihood estimation method. However, when the data has outlier from a single outlier to 30% outliers, the slope estimate and standard deviations using maximum likelihood estimation method becomes huge and break down completely. The proposed modified maximum likelihood estimation method is not affected by the increasing in the percentage of outliers.

3.7 Summary and Conclusions

This chapter proposes a modified maximum likelihood method based on the normality assumption for estimating the slope parameters of the unreplicated linear functional relationship model in the presence of outliers. In this model, the assumption is made on the ratio of error variances. The value of the ratio of error variances is equal to one, $\lambda = 1$ to overcome the unidentifiability problem. As the slope parameter in unreplicated linear functional relationship model is depending on the value of sample variances and the sample covariance which are known sensitive to outliers, the robust method is needed to overcome this problem. In the simulation study, several distributions for the error terms are considered, namely the symmetric and non-symmetric distributions. Different situations regarding the percentages of outliers in dataset also considered. The performance of the proposed method is evaluated using the estimated bias and the mean square error. In all the cases considered, the results of the simulation study provide numerical evidence of the advantages of the modified maximum likelihood estimation method when outliers are present in the data. For practical illustrations, two datasets are used using the two methods. From both simulation studies and examples, it can be

summarized that the maximum likelihood estimation method performs better when the data has no outlier. However, when there is an outlier present from as low as a single outlier to 30% outliers in the data, the maximum likelihood estimation method breaks down completely. On the other hand, the modified maximum likelihood estimation method performs very well in every cases. These results indicate that the proposed method is superior if there is a single outlier presence in the dataset. In conclusion, the proposed modified maximum likelihood estimation method is able to produce satisfactory results and provides a good alternative to the standard maximum likelihood estimator. Thus, the robust method called the modified maximum likelihood estimation method can be used to estimate the slope parameters of unreplicated linear functional relationship model in the presence of outliers. In general, the study has contributed to the body of knowledge on studies of parameter estimation of unreplicated linear functional relationship model. The implication of the finding is that a robust estimate of the parameter can be obtained without much mathematical complexity.

CHAPTER 4: PARAMETER ESTIMATION FOR REPLICATED LINEAR FUNCTIONAL RELATIONSHIP MODEL

4.1 Introduction

Unlike the previous chapter, this chapter discusses on another model namely the replicated linear functional relationship model where derivation of the estimation of parameters as well as the variance-covariance matrix is obtained. The replicated linear functional relationship model in this chapter is focused on balanced observations in each group and thus the scope of the study. The motivation of this study is that, the replicated linear functional relationship model can be used to overcome the unidentifiability problem in linear functional relationship model. Furthermore, the replicated linear functional relationship model can estimate all the parameters namely the intercept, the slope, the incidental parameters and also two error variances unlike unreplicated linear functional relationship model. The sections of the chapter are as follows. Derivation of maximum likelihood estimation of the replicated linear functional relationship model is described in Section 4.2. This is followed by the covariance matrix of the parameters for the replicated linear functional relationship model where the derivation is given in Section 4.3. Section 4.4 describes the simulation study for balanced replicated linear functional relationship model where the accuracy of the estimated parameters is investigated. Section 4.5 discusses on the result obtained from the simulation study. The applicability of the proposed method is illustrated by real data set in Section 4.6. Finally, summary and conclusion are provided in Section 4.7.

4.2 Maximum Likelihood Estimation Method

In Chapter 2, the parameters of linear functional relationship model have been discussed in general in terms of unreplicated and replicated linear functional relationship model. The parameters to be estimated in linear functional relationship model are the intercept, $\hat{\alpha}$, the slope, $\hat{\beta}$, the incidental parameter, $\hat{X}_i$, and two error variances, $\hat{\sigma}^2$ and $\hat{\tau}^2$ respectively. The maximum likelihood estimation (MLE) is often used in estimating the parameters because of the properties of maximum likelihood estimation method namely consistency, efficiency and normally distributed. Nevertheless, the unidentifiability problem become a major setback in estimating all the parameters in the linear functional relationship model.

In unreplicated linear functional relationship model, the unindentifibility problem can be avoided if there is a knowledge on ratio of error variances, namely $\lambda = \frac{\tau^2}{\sigma^2}$ is known in order to estimate the parameters (Lindley, 1947). However, this knowledge i.e. the value of λ is often unknown in most practical situations due to fact that the information is either not available or is not shared by the researchers in the field (Klepper & Leamer, 1984). In order to overcome this problem, one of practical approach is by obtaining this information from the sample itself. This can be done by considering the approach where the groups of observations can be identified from the unreplicated linear data where all parameters are identifiable and consistently estimated (Hussin, et al., 2005). Thus, in this chapter, it is discussed further the replicated linear functional relationship model by replicating observations either from unreplicated data or making replication when it is available. As mentioned earlier, the replicated linear functional relationship model can be used to overcome the unidentifiability problem in linear functional relationship model. Furthermore, by using replicated linear functional relationship model, the assumption or

the knowledge on the ratio of the error variances is no longer needed when one can estimate the error variances $\hat{\sigma}^2$ and $\hat{\tau}^2$ independently and then estimate other parameters by maximum likelihood estimation.

In this section, it is assumed the size of every group are the same which means that the observations or elements in each group are the same. This is called the equal and balanced replicates where measurements $x_{ij}, (j = 1,2, \dots, m)$ are made on X_i and measurements $y_{ik}, (k = 1,2, \dots, m)$ are made on Y_i are equal respectively i.e. $j = k = 1,2, \cdots, m$. Given a particular pair (X_i, Y_i) , there may be replicated observations of X_i and Y_i occurring in p groups. For this model, the linear relationship between X_i and Y_i are given by

$$x_{ij} = X_i + \delta_{ij} \text{ and } y_{ij} = Y_i + \varepsilon_{ij} \tag{4.1}$$

where $Y_i = \alpha + \beta X_i$ for $i = 1,2, \cdots, p$ and $j = 1,2, \cdots, m$ and $n = p \times m$.

The errors terms δ_{ij} and ε_{ij} follow normal distribution with mean zero and variance σ^2 and τ^2 respectively. This implies that

i) both errors have mean 0, that is $E(\delta_{ij}) = 0$ and $E(\varepsilon_{ij}) = 0$, $i = 1,2, \dots, p$ and $j = 1,2, \cdots, m$.

ii) both errors have constant but different variance, that is $Var(\delta_{ij}) = \sigma^2$ and $Var(\varepsilon_{ij}) = \tau^2$, $i = 1,2, \dots, p$ and $j = 1,2, \cdots, m$.

The estimation of parameter for balanced replicated linear functional relationship model can be obtained by maximum likelihood estimation method which involves an iterative technique as the closed-form solution is not available. In this case, the log likelihood function of balanced replicated linear functional relationship model can be expressed as

$$\log L\left(\alpha,\beta,\sigma^2,\tau^2,X_1,\dots,X_p;x_{ij},y_{ij}\right)=-n\log 2\pi-\frac{n}{2}(\log\sigma^2+\log\tau^2) \tag{4.2}$$

$$-\frac{1}{2}\left\{\sum_{i=1}^{p}\sum_{j=1}^{m}\frac{\left(x_{ij}-X_i\right)^2}{\sigma^2}+\sum_{i=1}^{p}\sum_{j=1}^{m}\frac{\left(y_{ij}-\alpha-\beta X_i\right)^2}{\tau^2}\right\}$$

There are $(p+4)$ parameters to be estimated as given in (4.2) namely the intercept, α, the slope, β, the error variances σ^2 and $\tau^2$and also the incidental parameters, X_i. Thus, to estimate $\hat{\alpha},\hat{\beta},\hat{\sigma}^2,\hat{\tau}^2$ and $\hat{X}_i$, the first partial derivative of the log likelihood function with respect to $\alpha,\beta,\sigma^2,\tau^2$ and X_i is obtained. The subsequent sections described the maximum likelihood estimation method for all parameters e $\hat{\alpha},\hat{\beta},\hat{\sigma}^2,\hat{\tau}^2$ and $\hat{X}_i$.

4.2.1 Maximum Likelihood for α

The first partial derivative of the log likelihood function with respect to α is,

$$\frac{\partial logL}{\partial\alpha}=\frac{1}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij}-\alpha-\beta X_i\right) \tag{4.3}$$

By setting $\frac{\partial logL}{\partial \alpha} = 0$ of the log likelihood function, then,

$$\frac{1}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}(y_{ij} - \alpha - \beta X_i) = 0$$

Hence, expanding the equation above,

$$\sum_{i=1}^{p}\sum_{j=1}^{m} y_{ij} - \sum_{i=1}^{p}\sum_{j=1}^{m} \alpha - \sum_{i=1}^{p}\sum_{j=1}^{m} \beta X_i = 0$$

$$\sum_{i=1}^{p}\sum_{j=1}^{m} \alpha = \sum_{i=1}^{p}\sum_{j=1}^{m} y_{ij} - \sum_{i=1}^{p}\sum_{j=1}^{m} \beta X_i$$

Upon simplifying, it is then given by,

$$\hat{\alpha} = \frac{\sum_{i=1}^{p} m(\bar{y}_{i.} - \hat{\beta}\hat{X}_i)}{\sum_{i=1}^{p} m} \tag{4.4}$$

where $\bar{y}_{i.} = \frac{\sum_{j=1}^{m} y_{ij}}{m}$.

4.2.2 Maximum Likelihood for $\boldsymbol{\beta}$

The first partial derivative of the log likelihood function with respect to β is

$$\frac{\partial logL}{\partial \beta} = \frac{X_i}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}(y_{ij} - \alpha - \beta X_i) \tag{4.5}$$

By setting $\frac{\partial logL}{\partial \beta} = 0$,

$$\frac{X_i}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}(y_{ij}-\alpha-\beta X_i)=0$$

Hence, expanding the equation above,

$$\sum_{i=1}^{p}\sum_{j=1}^{m}X_i y_{ij}-\sum_{i=1}^{p}\sum_{j=1}^{m}X_i\alpha-\sum_{i=1}^{p}\sum_{j=1}^{m}\beta X_i^2=0$$

$$\sum_{i=1}^{p}\sum_{j=1}^{m}\beta X_i^2=\sum_{i=1}^{p}\sum_{j=1}^{m}X_i y_{ij}-\sum_{i=1}^{p}\sum_{j=1}^{m}X_i\alpha$$

Simplify to get,

$$\hat{\beta}=\frac{\sum_{i=1}^{p}m\hat{X}_i(\bar{y}_{i.}-\hat{\alpha})}{\sum_{i=1}^{p}m\hat{X}_i^2} \tag{4.6}$$

where $\bar{y}_{i.}=\frac{\sum_{j=1}^{m}y_{ij}}{m}$.

4.2.3 Maximum Likelihood for σ^2

The first partial derivative of the log likelihood function with respect to σ^2 is

$$\frac{\partial logL}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{\sum_{i=1}^{p}\sum_{j=1}^{m}(x_{ij}-X_i)^2}{2\sigma^4} \tag{4.7}$$

where $n=\sum_{i=1}^{p}m=m\times p$.

By setting $\frac{\partial logL}{\partial\sigma^2}=0$, the equation can be written as,

$$-\frac{n}{2\sigma^2}+\frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(x_{ij}-X_i\right)^2}{2\sigma^4}=0$$

$$\frac{n}{2\sigma^2}=\frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(x_{ij}-X_i\right)^2}{2\sigma^4}$$

$$\frac{2\sigma^4}{2\sigma^2}=\frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(x_{ij}-X_i\right)^2}{n}$$

Simplify to get,

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(x_{ij}-\hat{X}_i\right)^2}{\sum_{i=1}^{p}m} \tag{4.8}$$

where $\bar{x}_{i.}=\frac{\sum_{j=1}^{m}x_{ij}}{m}$ and $n=\sum_{i=1}^{p}m=p\times m$.

4.2.4 Maximum Likelihood for τ^2

The first partial derivative of the log likelihood function with respect τ^2 is

$$\frac{\partial logL}{\partial\tau^2}=-\frac{n}{2\tau^2}+\frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij}-\alpha-\beta X_i\right)^2}{2\tau^4} \tag{4.9}$$

By setting $\frac{\partial logL}{\partial\tau^2}=0$, the equation can be written as,

$$-\frac{n}{2\tau^2}+\frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij}-\alpha-\beta X_i\right)^2}{2\tau^4}=0$$

$$\frac{n}{2\tau^2}=\frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij}-\alpha-\beta X_i\right)^2}{2\tau^4}$$

$$\frac{2\tau^4}{2\tau^2} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij} - \alpha - \beta X_i\right)^2}{n}$$

Simplify to get

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij} - \hat{\alpha} - \hat{\beta}\hat{X}_i\right)^2}{\sum_{i=1}^{p} m} \tag{4.10}$$

where $\bar{y}_{i.} = \frac{\sum_{j=1}^{m} y_{ij}}{m}$ and $n = \sum_{i=1}^{p} m = p \times m$.

4.2.5 Maximum Likelihood for X_i

The first partial derivative of the log likelihood function with respect X_i is

$$\frac{\partial logL}{\partial X_i} = \frac{1}{\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{m}\left(x_{ij} - X_i\right) + \frac{\beta}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij} - \alpha - \beta X_i\right) \tag{4.11}$$

By setting $\frac{\partial logL}{\partial X_i} = 0$, the equation can be written as,

$$\frac{1}{\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{m}\left(x_{ij} - X_i\right) + \frac{\beta}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij} - \alpha - \beta X_i\right) = 0$$

Upon simplifying, it is then given by,

$$\hat{X}_i = \frac{1}{\hat{\Delta}}\left\{\frac{m\bar{x}_{i.}}{\hat{\sigma}^2} + \frac{m\hat{\beta}}{\hat{\tau}^2}\left(\bar{y}_{i.} - \hat{\alpha}\right)\right\} \tag{4.12}$$

where $\bar{x}_{i.} = \frac{\sum x_{ij}}{m}$, $\bar{y}_{i.} = \frac{\sum y_{ij}}{m}$, and $\hat{\Delta}_i = \frac{m}{\hat{\sigma}^2} + \frac{m\hat{\beta}^2}{\hat{\tau}^2}$

From equations (4.4), (4.6), (4.8) and (4.10) , the parameter intercept, α, the slope, β, the error variances σ^2 and τ^2 are depending on the value of $\hat{X}_i$ as given in (4.12) which clearly shows there is no closed form available. Hence, to get the solution for parameter α, β, σ^2 and τ^2, an iteration procedure is needed and initial value for each parameters are required. As for the estimation $\hat{X}_i$, the initial value $\alpha_0, \beta_0, \sigma_0^2$ and τ_0^2 can be obtained from unreplicated linear functional relationship model with the assumption ratio of error variances are equal one, $\lambda = \frac{\tau_0^2}{\sigma_0^2} = 1$ or $\sigma_0^2 = \tau_0^2$ to start the iteration process. This will become,

$$\hat{X}_i = \frac{1}{\hat{\Delta}}\left\{\frac{m\bar{x}_{i.}}{\hat{\sigma}_0^2} + \frac{m\hat{\beta}_0}{\hat{\tau}_0^2}(\bar{y}_{i.} - \hat{\alpha}_0)\right\} \tag{4.13}$$

4.3 Fisher Information Matrix -Variance Covariance Matrix

In this section, the asymptotic variances of the estimators are derived by inverting the estimated Fisher information matrix for equal and balanced replicated linear functional relationship model. From previous section, the first partial derivatives of log likelihood function of *L* with respect to $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2$ and $\hat{X}_i$ have been derived from (4.3), (4.5), (4.7), (4.9) and (4.11) respectively. The second partial derivatives for log likelihood function and their negative expected values are given by equations (4.14) up to (4.24):

$$\frac{\partial^2 logL}{\partial \alpha^2} = -\frac{1}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}(1)\text{, hence } E\left[-\frac{\partial^2 logL}{\partial \alpha^2}\right] = \frac{mp}{\tau^2} \tag{4.14}$$

$$\frac{\partial^2 logL}{\partial \beta^2} = -\frac{1}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}X_i^2\text{, hence } E\left[-\frac{\partial^2 logL}{\partial \beta^2}\right] = \frac{m\sum_{i=1}^{p}X_i^2}{\tau^2} \tag{4.15}$$

$$\frac{\partial^2 logL}{\partial \alpha \partial \beta} = -\frac{1}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}X_i\text{, hence } E\left[-\frac{\partial^2 logL}{\partial \alpha \partial \beta}\right] = \frac{m\sum_{i=1}^{p}X_i}{\tau^2} \tag{4.16}$$

$$\frac{\partial^2 logL}{\partial (\sigma^2)^2} = \frac{\sum_{i=1}^{p} m}{2\sigma^4} - \frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(x_{ij} - X_i\right)^2}{\sigma^6} \tag{4.17}$$

where $n = \sum_{i=1}^{p} m = m \times p$ and hence,

$$E\left[-\frac{\partial^2 logL}{\partial (\sigma^2)^2}\right] = E\left[-\frac{\sum_{i=1}^{p} m}{2\sigma^4} + \frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(x_{ij} - X_i\right)^2}{\sigma^6}\right]$$

$$= -\frac{\sum_{i=1}^{p} m}{2\sigma^4} + \frac{\sum_{i=1}^{p} m}{\sigma^4} = \frac{mp}{2\sigma^4} = \frac{n}{2\sigma^4}$$

$$\frac{\partial^2 logL}{\partial (\tau^2)^2} = \frac{\sum_{i=1}^{p} m}{2\tau^4} - \frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij} - \alpha - \beta X_i\right)^2}{2\tau^6}, \tag{4.18}$$

where $n = \sum_{i=1}^{p} m = m \times p$ and hence,

$$E\left[-\frac{\partial^2 logL}{\partial (\tau^2)^2}\right] = \left[-\frac{\sum_{i=1}^{p} m}{2\tau^4} + \frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij} - \alpha - \beta X_i\right)^2}{2\tau^6}\right]$$

$$= -\frac{\sum_{i=1}^{p} m}{2\tau^4} + \frac{\sum_{i=1}^{p} m}{\tau^4} = \frac{mp}{2\tau^4} = \frac{n}{2\tau^4}$$

$$\frac{\partial^2 logL}{\partial X_i^2} = -\frac{1}{\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{m}(1) - \frac{\beta^2}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{m}(1), \text{hence} \tag{4.19}$$

$$E\left[-\frac{\partial^2 logL}{\partial X_i^2}\right] = \frac{m}{\sigma^2} + \frac{m\beta^2}{\tau^2}$$

$$\frac{\partial^2 logL}{\partial X_i X_j} = 0, \text{hence } E\left[-\frac{\partial^2 logL}{\partial X_i X_j}\right] = 0 \tag{4.20}$$

$$\frac{\partial^2 logL}{\partial X_i \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^{p}\sum_{j=1}^{m}(x_{ij} - X_i), \text{hence } E\left[-\frac{\partial^2 logL}{\partial X_i \partial \sigma^2}\right] = 0 \tag{4.21}$$

$$\frac{\partial^2 logL}{\partial X_i \partial \tau^2} = -\frac{1}{\tau^4}\sum_{i=1}^{p}\sum_{j=1}^{m}(y_{ij} - \alpha - \beta X_i), \text{hence } E\left[-\frac{\partial^2 logL}{\partial X_i \partial \tau^2}\right] = 0 \tag{4.22}$$

$$\frac{\partial^2 logL}{\partial X_i \partial \alpha} = -\frac{m\beta}{\tau^2}, \text{hence } E\left[-\frac{\partial^2 logL}{\partial X_i \partial \alpha}\right] = \frac{m\beta}{\tau^2} \tag{4.23}$$

$$\frac{\partial^2 logL}{\partial X_i \partial \beta} = -\frac{m\beta\sum_{i=1}^{p} X_i}{\tau^2}, \text{hence } E\left[-\frac{\partial^2 logL}{\partial X_i \partial \beta}\right] = \frac{m\beta\sum_{i=1}^{p} X_i}{\tau^2} \tag{4.24}$$

Next, the estimated Fisher information matrix, F, for $\hat{X}_i, \cdots, \hat{X}_i, \hat{\sigma}^2, \hat{\tau}^2$, $\hat{\alpha}$ and $\hat{\beta}$ can be obtained and this is given by

$$F = \begin{bmatrix} B & 0 & E \\ 0 & C & 0 \\ E^T & 0 & D \end{bmatrix} \tag{4.25}$$

where B is a $p \times p$ matrix given by

$$B = \begin{bmatrix} \frac{m}{\sigma^2} + \frac{m\beta^2}{\tau^2} & & 0 \\ & \ddots & \\ 0 & & \frac{m}{\sigma^2} + \frac{m\beta^2}{\tau^2} \end{bmatrix}$$

E is a $p \times 2$ matrix given by

$$E = \begin{bmatrix} \frac{m\beta}{\tau^2} & \frac{mX_1\beta}{\tau^2} \\ \vdots & \vdots \\ \frac{m\beta}{\tau^2} & \frac{mX_p\beta}{\tau^2} \end{bmatrix}$$

C is a 2×2 matrix given by

$$C = \begin{bmatrix} \frac{n}{2\sigma^4} & 0 \\ 0 & \frac{n}{2\tau^4} \end{bmatrix}$$

D is a 2×2 matrix given by

$$D = \begin{bmatrix} \frac{mp}{\tau^2} & \frac{m\sum_{i=1}^{p} X_i}{\tau^2} \\ \frac{m\sum_{i=1}^{p} X_i}{\tau^2} & \frac{m\sum_{i=1}^{p} X_i^2}{\tau^2} \end{bmatrix} = \frac{m}{\tau^2}\begin{bmatrix} p & \sum_{i=1}^{p} X_i \\ \sum_{i=1}^{p} X_i & \sum_{i=1}^{p} X_i^2 \end{bmatrix}$$

The primary interest is the bottom right minor of order $4 \times p$ of the inverse of matrix F as in (4.25) where the value of the of $\hat{\alpha}$, $\hat{\beta}$,$\hat{\sigma}^2$ and $\hat{\tau}^2$ is positioned. Thus, the asymptotic

covariance matrix of $\hat{\sigma}^2, \hat{\tau}^2, \hat{\alpha}$ and $\hat{\beta}$ can be obtained using the theory of partitioned matrices (Graybill, 1961). Details on the asymptotic covariance matrix is given in APPENDIX E. The covariance matrix is given by:

$$\widehat{Var}\begin{bmatrix}\hat{\sigma}^2\\ \hat{\tau}^2\\ \hat{\alpha}\\ \hat{\beta}\end{bmatrix} = \begin{bmatrix} C^{-1} & 0 \\ 0 & (D - E^T B^{-1} E)^{-1}\end{bmatrix}$$

where $C^{-1} = \begin{bmatrix} 2\sigma^4/n & 0 \\ 0 & 2\tau^4/n \end{bmatrix}$ and

$$(D - E^T B^{-1} E)^{-1} = \frac{m\tau^2 + m\beta^2\sigma^2}{m^2\left\{p\sum_{i=1}^{p} X_i^2 - \left(\sum_{i=1}^{p} X_i\right)^2\right\}}\begin{bmatrix} \sum_{i=1}^{p} X_i^2 & -\sum_{i=1}^{p} X_i \\ -\sum_{i=1}^{p} X_i & p \end{bmatrix}$$

Therefore, the asymptotic covariance matrix for $\hat{\sigma}^2, \hat{\tau}^2, \hat{\alpha}$ and $\hat{\beta}$ are given by

$$G = \begin{bmatrix} 2\sigma^4/n & 0 & 0 & 0 \\ 0 & 2\tau^4/n & 0 & 0 \\ 0 & 0 & Q\sum_{i=1}^{p} X_i^2 & -Q\sum_{i=1}^{p} X_i \\ 0 & 0 & -Q\sum_{i=1}^{p} X_i & Qp \end{bmatrix} \tag{4.26}$$

where

$$Q = \frac{m\tau^2 + m\beta^2\sigma^2}{m^2\left\{p\sum_{i=1}^{p} X_i^2 - \left(\sum_{i=1}^{p} X_i\right)^2\right\}}$$

By using Fisher Information matrix, the estimated covariance of parameters is derived. In particular, the following results are given as follows:

$$\widehat{Var}\,(\hat{\sigma}^2) = 2\sigma^4/n \tag{4.27}$$

$$\widehat{Var}\,(\hat{\tau}^2) = 2\tau^4/n \tag{4.28}$$

$$\widehat{Var}\,(\hat{\alpha}) = \frac{(m\tau^2 + m\beta^2\sigma^2)\sum_{i=1}^{p} X_i^2}{m^2\left\{p\sum_{i=1}^{p} X_i^2 - \left(\sum_{i=1}^{p} X_i\right)^2\right\}} \tag{4.29}$$

$$\widehat{Var}\,(\hat{\beta}) = \frac{(m\tau^2 + m\beta^2\sigma^2)p}{m^2\left\{p\sum_{i=1}^{p} X_i^2 - \left(\sum_{i=1}^{p} X_i\right)^2\right\}} \tag{4.30}$$

$$\widehat{Cov}\,(\hat{\alpha},\hat{\beta}) = -\frac{(m\tau^2 + m\beta^2\sigma^2)\sum_{i=1}^{p} X_i}{m^2\left\{p\sum_{i=1}^{p} X_i^2 - \left(\sum_{i=1}^{p} X_i\right)^2\right\}} \tag{4.31}$$

4.4 Simulation Studies

A simulation study is carried out to support the algebraic results presented earlier by assessing the accuracy and measuring the biasness of the parameters of the proposed model. The computer program (APPENDIX F) is written in R to carry out the simulation study. The value of error variances and the sample sizes are the two aspects to be considered for performance. Without loss of generality, the study considers the parameter

settings $\alpha = 0$ and different estimated values of $\beta = 0.8, 1, 1.2$, $\sigma^2 = 0.8, 1$ and $\tau^2 = 0.8,1$. For each specified set of parameter values, 10000 simulated data sets are obtained for each of sample sizes $n = 20, 50, 100, 180$ and 300 respectively. These sample sizes are chosen to represent both small and large datasets.

The simulation is considered for balanced replicates of the data in which data of x_{ij} and y_{ij} variables are of equal sample size. Furthermore, the method of grouping the observations in general is proposed. The procedure in arranging and grouping the data for the simulation study can be described in the following steps:

Step 1: Generate $X_i = 10\left(\frac{i}{p}\right)$ of size p, with $i = 1,2, \dots, p$ where p is the number of group and two error terms δ_{ij} and ε_{ij} from Normal distribution with mean 0 and variances σ^2 and τ^2 respectively; $N(0, \sigma^2)$ and $N(0, \tau^2)$ with $j = 1,2, \dots, m$ where m is the number of elements in each subgroups and m is not necessarily less than p.

Step 2: Calculate the observed values of x_{ij} and y_{ij} as in (4.1).

Step 3: The values of x_{ij} and y_{ij} are divided into p-subgroups with m elements such that $p \times m = n$ to obtain the groups of the data as given in Table 4.1.

Step 4: In this case, all parameters namely $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2$ and $\hat{X}_i$ can be solved iteratively given some suitable initial values at the estimate. The parameter X_i as in (4.12) need to be estimated first due to fact that other parameters $\alpha, \beta, \sigma^2, \tau^2$ as in (4.4), (4.6), (4.8), (4.10) respectively are dependent on the value of X_i. The possible initial estimates for $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$ and $\hat{\tau}^2$ can be obtained from unreplicated linear functional relationship model ($\alpha_0, \beta_0, \sigma_0^2$ and τ_0^2) with the assumption ratio of error variances are equal to one ($\lambda = 1$) to start

the iteration process. The final estimation for all parameters, $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2$ and $\hat{X}_i$ will be obtained after one or all the values have been converged.

Step 5: Step 1- Step 4 are repeated for 10000 simulations.

Table 4.1 The division of the sample size into their subgroup

Sample size, n	Subgroups, p	Number of elements, m
20	4	5
50	5	10
100	10	10
180	12	15
300	15	20

Table 4.1 shows some examples of how many the observations are grouped together to obtain the group of data. The number of observations or the elements are equal in each group. The performance of the estimated parameters is calculated using the estimated bias (EB), the mean square error (MSE) and the standard deviation (SD). The estimated bias and the mean square error are defined by:

$$\text{Estimated Bias, EB} = |\widehat{w} - w| \text{ and} \tag{4.32}$$

$$\text{Mean Square Error, MSE} = \frac{1}{s}\sum\left(\widehat{w}_j - w\right)^2 \tag{4.33}$$

with *w* be a generic term for the parameters and *s* is the number of simulation. The standard deviation (SD) for each parameter are computed by taking square root from the diagonal element of the asymptotic variance-covariance matrix (4.26) or from equation (4.27) to (4.30).

4.5 Results and Discussion

The results from simulation study are tabulated in Table 4.2, Table 4.3, Table 4.4 and Table 4.5 respectively with different values of error variances and the sample sizes.

Table 4.2 Parameter Estimates when $\alpha = 0, \beta = 1, \sigma^2 = 1$ and $\tau^2 = 1$

Statistic	Sample size	Parameter Estimates			
		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\tau}^2$
Estimated Bias	20	0.0321	0.0060	0.1397	0.1495
	50	0.0100	0.0018	0.0655	0.0689
	100	0.0101	0.0019	0.0590	0.0567
	180	0.0015	0.0004	0.0364	0.0393
	300	0.0015	0.0002	0.0278	0.0275
Mean Square Error	20	0.6565	0.0140	0.1168	0.1160
	50	0.2325	0.0053	0.0434	0.0429
	100	0.0976	0.0025	0.0234	0.0233
	180	0.0525	0.0014	0.0123	0.0127
	300	0.0303	0.0008	0.0075	0.0073
Standard Deviation	20	0.7147	0.1044	0.2721	0.2689
	50	0.4523	0.0682	0.1869	0.1862
	100	0.2962	0.0477	0.1331	0.1334
	180	0.2201	0.0359	0.1016	0.1013
	300	0.1693	0.0279	0.0794	0.0794

For each specified set of parameters in Table 4.2, for the error variances $\sigma^2 = \tau^2$, the estimated bias of parameters $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$ and $\hat{\tau}^2$ are consistently small when the sample size increase from 20 to 300. The value of the estimated bias approximately close to 0 shows the unbiasedness of the estimated parameters. The mean square error also shows similar trends such as these values tends to decrease with the increase in sample sizes. This shows that the estimated values of parameters are consistent. Moreover, the standard deviation is generally small for all parameter estimates.

Table 4.3 Parameter Estimates when $\alpha = 0, \beta = 0.8, \sigma^2 = 1$ and $\tau^2 = 0.8$

Statistic	Sample size	Parameter Estimates			
		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\tau}^2$
Estimated Bias	20	0.0335	0.0060	0.1448	0.1145
	50	0.0105	0.0022	0.0687	0.0527
	100	0.0072	0.0012	0.0624	0.0423
	180	0.0002	0.0002	0.0394	0.0280
	300	0.0001	0.0001	0.0311	0.0193
Mean Square Error	20	0.4618	0.0099	0.1175	0.0722
	50	0.1599	0.0036	0.0435	0.0270
	100	0.0706	0.0018	0.0238	0.0145
	180	0.0373	0.0010	0.0126	0.0079
	300	0.0209	0.0006	0.0077	0.0046
Standard Deviation	20	0.6110	0.0893	0.2704	0.2168
	50	0.3847	0.0580	0.1863	0.1495
	100	0.2515	0.0405	0.1326	0.1072
	180	0.1869	0.0305	0.1013	0.0814
	300	0.1436	0.0237	0.0791	0.0637

By looking at Table 4.3 when $\beta = 0.8$ and $\sigma^2 > \tau^2$, it is observed the value of estimated bias is approximately close to 0 which means the value of the estimated

parameters is approximately close to the true mean. The mean square error for the estimated parameters are also observed to decrease with the increasing of sample sizes from 20 to 300. This shows the estimated parameters are consistent. The same trend can be said for the standard deviation.

Table 4.4 Parameter Estimates when $\alpha = 0, \beta = 0.8, \sigma^2 = 0.8$ and $\tau^2 = 1$

Statistic	Sample size	Parameter Estimates			
		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\tau}^2$
Estimated Bias	20	0.0321	0.0060	0.1397	0.1495
	50	0.0100	0.0018	0.0655	0.0689
	100	0.0101	0.0019	0.0590	0.0567
	180	0.0015	0.0004	0.0364	0.0393
	300	0.0015	0.0002	0.0278	0.0275
Mean Square Error	20	0.6565	0.0140	0.1168	0.1160
	50	0.2325	0.0053	0.0434	0.0429
	100	0.0976	0.0025	0.0234	0.0233
	180	0.0525	0.0014	0.0123	0.0127
	300	0.0303	0.0008	0.0075	0.0073
Standard Deviation	20	0.7147	0.1044	0.2721	0.2689
	50	0.4523	0.0682	0.1869	0.1862
	100	0.2962	0.0477	0.1331	0.1334
	180	0.2201	0.0359	0.1016	0.1013
	300	0.1693	0.0279	0.0794	0.0794

Table 4.5 Parameter Estimates when $\alpha = 0, \beta = 1.2, \sigma^2 = 0.8$ and $\tau^2 = 1$

Statistic	Sample size	Parameter Estimates			
		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\tau}^2$
Estimated Bias	20	0.0405	0.0072	0.1088	0.1524
	50	0.0146	0.0029	0.0504	0.0715
	100	0.0104	0.0017	0.0436	0.0608
	180	0.0018	0.0006	0.0272	0.0403
	300	0.0017	0.0004	0.0216	0.0282
Mean Square Error	20	0.6849	0.0147	0.0739	0.1149
	50	0.2383	0.0054	0.0276	0.0428
	100	0.1049	0.0027	0.0148	0.0234
	180	0.0556	0.0015	0.0078	0.0126
	300	0.0312	0.0009	0.0048	0.0073
Standard Deviation	20	0.7461	0.1091	0.2186	0.2680
	50	0.4704	0.0709	0.1499	0.1857
	100	0.3076	0.0496	0.1070	0.1328
	180	0.2286	0.0373	0.0815	0.1012
	300	0.1757	0.0290	0.0636	0.0793

Table 4.4 and Table 4.5 show the results for difference value of slope, β when $\sigma^2 < \tau^2$. Similar results can be obtained for estimated bias, mean square error and standard deviation for the estimated parameters. As the sample size increase from 20 to 300, the values of estimated bias, the mean square error and the standard deviation also decrease.

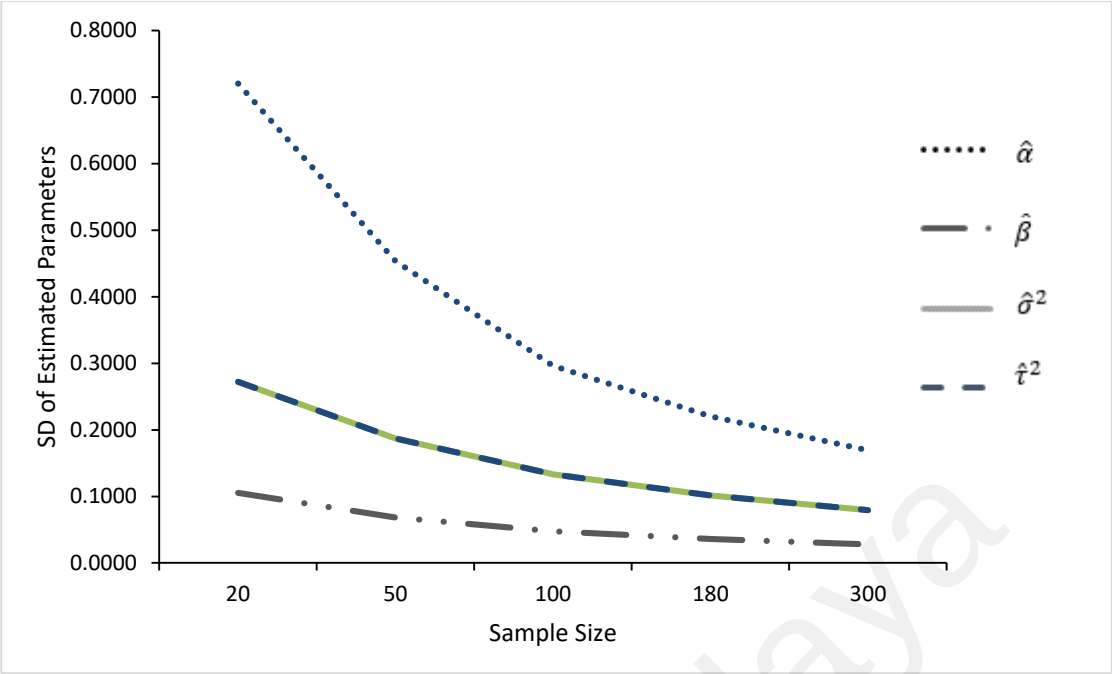

Figure 4.1 Standard Deviations for parameter $\alpha = 0, \beta = 1, \sigma^2 = 1$ and $\tau^2 = 1$

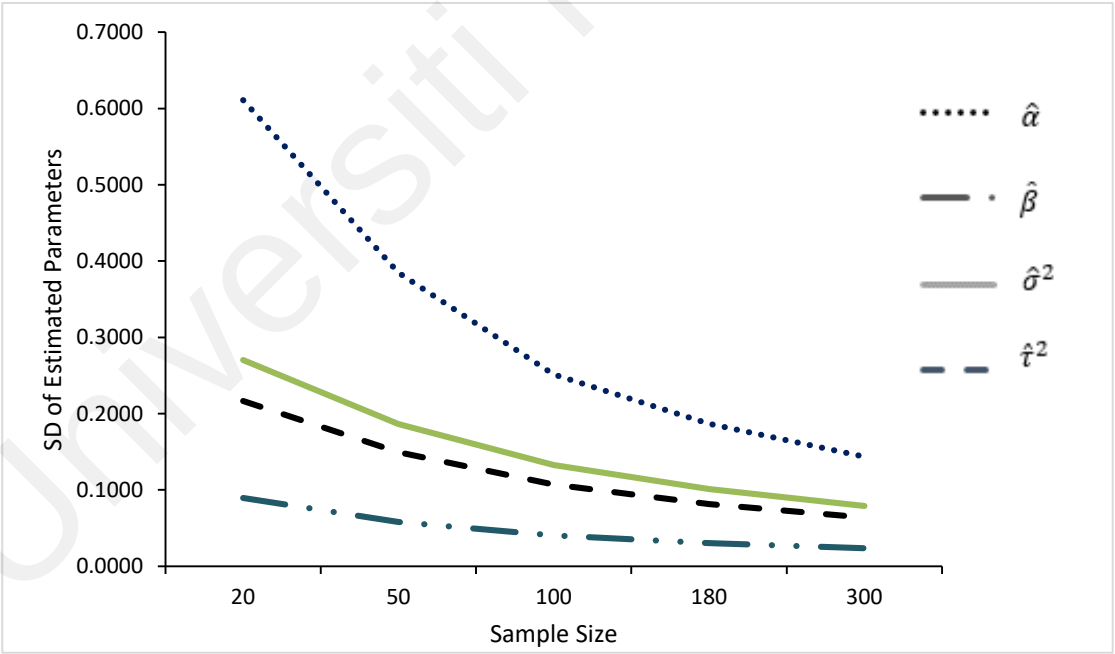

Figure 4.2 Standard Deviations for parameters $\alpha = 0, \beta = 0.8, \sigma^2 = 1$ and $\tau^2 = 0.8$

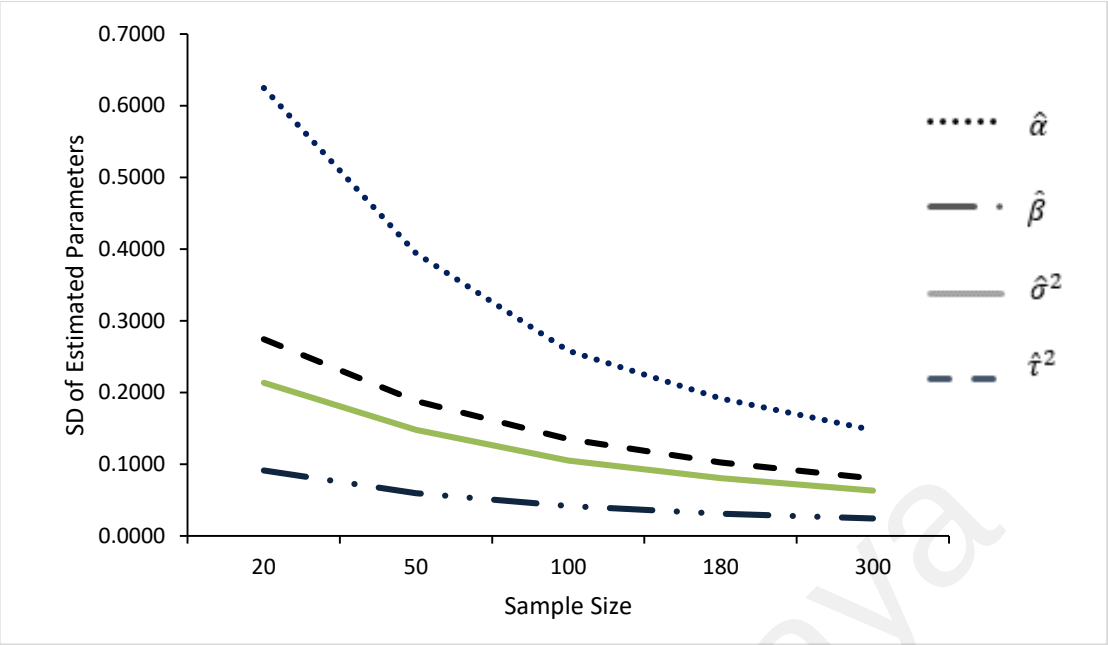

Figure 4.3 Standard Deviations for parameters $\alpha = 0, \beta = 0.8, \sigma^2 = 0.8$ and $\tau^2 = 1$

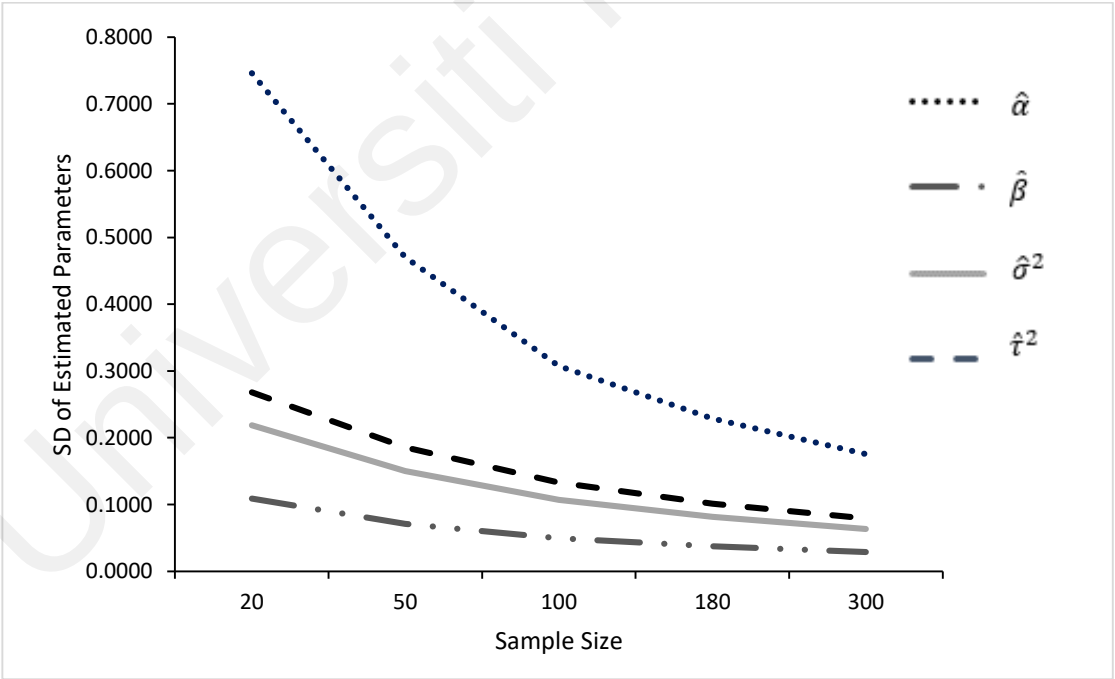

Figure 4.4 Standard Deviations for parameters $\alpha = 0, \beta = 1.2, \sigma^2 = 0.8$ and $\tau^2 = 1$

To illustrate the results obtained, plots of the standard deviations of different sets of estimated parameters for different sample sizes are given in Figure 4.1 to Figure 4.4. From these plots, the standard deviations of the estimators tend to decrease with the increase in sample sizes and number of replication and which also suggest that all the estimated parameters are consistent.

4.6 Examples

In this section, the applicability of the replicated linear functional relationship model when the observations are equal and balanced in each group is illustrated using two examples. The fat mass measurements data taken from Goran et al. (1996) and also systolic blood data taken from Altman and Bland (1999). It is assumed the measurement error can occur on these two examples in order to apply the relationship as in model **(**4.1).

4.6.1 Fat Mass Measurements Data

The dataset from Goran et. al (1996) consists of 96 observations taken on the children by using two techniques namely the skinfold thickness (ST), x_i and bioelectrical resistance (BR), y_i. This dataset can be considered as unreplicated data because there is only a single x and y observation for each level of i (Hussin et al., 2005). In unreplicated linear functional relationship model, one need an assumption on ratio of error variances, λ, to estimate the parameters namely the intercept, α, the slope, β and the error variance, σ^2. However, in the absence of knowledge on ratio of error variances, it is

important to transform the data into several groups and use maximum likelihood estimation method for balanced replicated linear functional relationship model to estimate all parameters namely the intercept, α, the slope, β and two error variances, σ^2 and also τ^2respectively.

Since there are 96 observations in this data, the group of the data is obtained by dividing the data into 8 groups and each groups have 12 observations that are balanced and equal in each group. The measurement X_i and Y_i are referring to skinfold thickness and bioelectrical resistance techniques respectively which are observed with errors. The data can be modelled by

$$Y_i = \alpha + \beta X_i$$

for $i = 1,2,\cdots,8$ and $j = 1,2,\cdots,12$ and $n = p \times m = 8 \times 12 = 96$. Figure 4.5 shows the scatterplot for the data. From this plot, there exists a linear relationship between two techniques i.e., the skinfold thickness (ST) and bioelectrical resistance (BR).

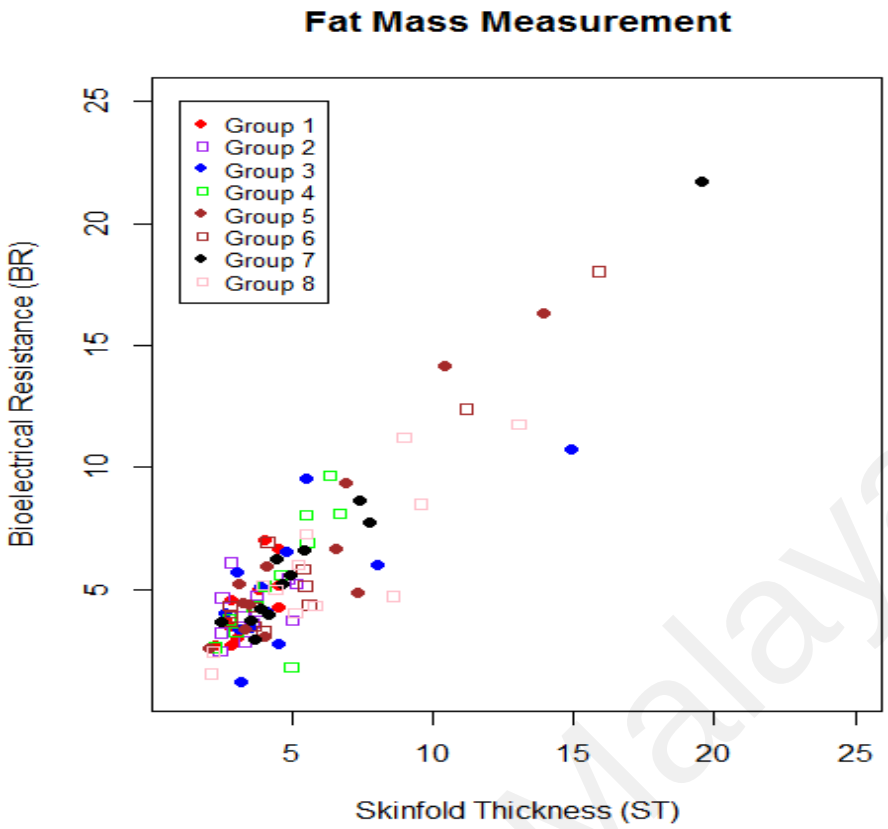

Figure 4.5 The scatterplot of Fat Mass Measurements Data

In order to investigate the relationship between two techniques, the data is fitted with balanced replicated linear functional relationship model. This is given by

$$Y_i = 0.91 + 0.93\, X_i \quad \text{for } i = 1,2,\cdots,8 \text{ and } j = 1,2,\cdots,12.$$

The estimated parameters and the standard deviations can be seen from Table 4.6. From this table, the value of the error variances of x_{ij} and y_{ij} are 8.58 and 10.53 respectively. It shows that the ratio of error variance is approximately 1.2. Furthermore, the estimated standard deviation for error variance parameter of measurement error by skinfold

thickness is less than the estimated standard deviation for error variance parameters by bioelectrical thickness which suggests the measurement error by skinfold thickness technique seems to be more precise.

Table 4.6 Estimated Parameters and Standard Deviations for fat mass measurements data

Parameter	Balanced Replicated	Standard Deviation
$\hat{\alpha}$	0.91	2.22
$\hat{\beta}$	0.93	0.44
$\hat{\sigma}^2$	8.58	1.24
$\hat{\tau}^2$	10.53	1.52

4.6.2 Systolic Blood Pressure Data

Next, another dataset taken from Altman and Bland (1999) is considered. Details on the dataset is given in APPENDIX C. Subsample of the original data containing 30 observations is used. The dataset measures the systolic blood pressure which simultaneous measurements were made by two experienced observers denoted as J (x_{ij}) and R (y_{ij}) . This data set can be considered as replicated data as there have 10 groups (or subjects) and each groups have three sets of readings that were made in quick succession. It is assumed that measurement error can occur in both the variables Y_i and X_i. The data can be modelled by

$$Y_i = \alpha + \beta X_i$$

for $i = 1,2,\cdots,10$ and $j = 1,2,3$ and $n = p \times m = 10 \times 3 = 30$.The scatterplot of the data can be seen in Figure 4.6.

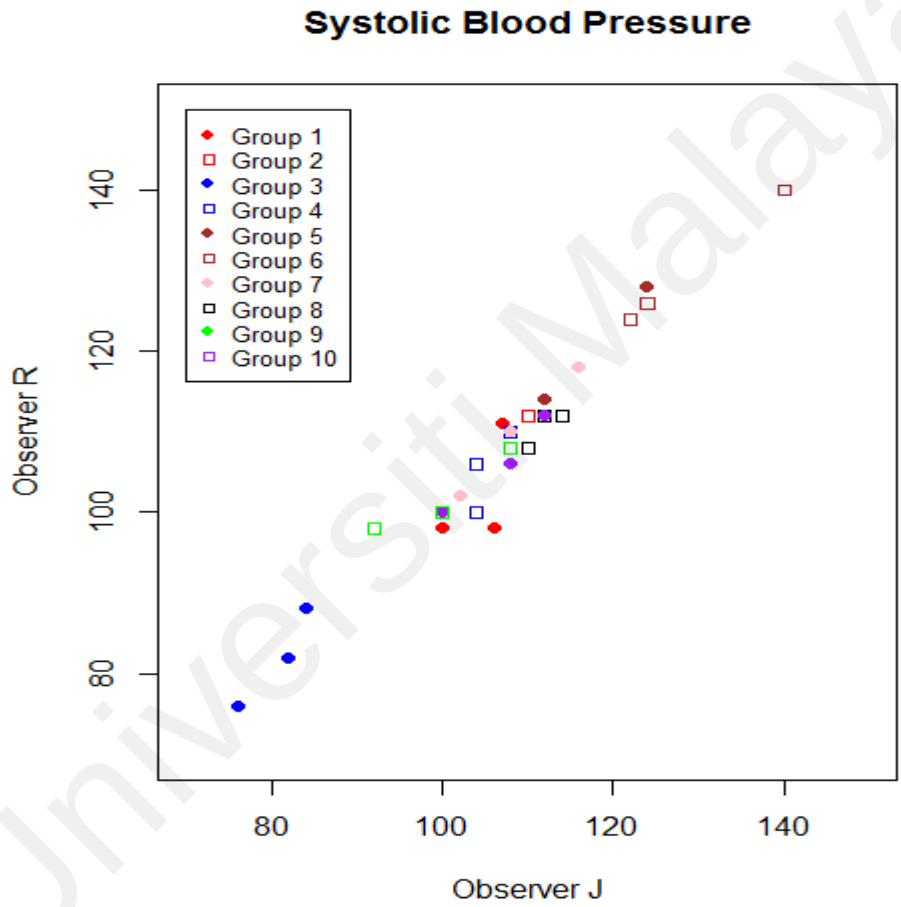

Figure 4.6 The scatterplot of Systolic Blood Pressure Data

In order to investigate the relationship between two experienced observers, the data is fitted with balanced replicated linear functional relationship model and is given by:

$$\hat{Y}_i = -0.81 + 1.01\ \hat{X}_i \quad \text{for } i = 1,2,\cdots,10 \text{ and } j = 1,2,3.$$

The values of the parameter estimate and their standard deviation are given in Table 4.7. The ratio of error variances is 1.2 with the value of error variance is 22.94 and 27.68 respectively. Moreover, the estimated standard deviation for error variance parameter of measurement error by observer J is less than the estimated standard deviation for error variance parameters by observer R which suggest the measurement error by observer J seems to be more precise.

Table 4.7 Estimated Parameters and Standard Deviations for systolic blood pressure data

Parameter	Balanced Replicated	Standard Deviation
$\hat{\alpha}$	-0.81	12.24
$\hat{\beta}$	1.01	0.11
$\hat{\sigma}^2$	22.94	5.92
$\hat{\tau}^2$	27.68	7.15

4.7 Summary and Conclusions

In this chapter, a balanced replicated linear functional relationship model for balanced and equal observations is proposed. The motivation of the model is that using replicated

linear functional relationship model, one can estimate all the parameters namely the intercept, the slope, the incidental parameters and also two error variances which cannot be done for unreplicated linear functional relationship model. In this chapter, the maximum likelihood estimation of parameters and the covariance matrix are derived. Although the closed form solution is not available in estimating the parameters, the estimated parameters can be obtained using iteration procedures and by choosing the suitable initial values. Nevertheless, the variance-covariance matrix can be obtained using the Fisher information matrix and partitioned matrix. Unlike the unreplicated linear functional relationship model, the replicated linear functional relationship model can estimate all the parameters using the maximum likelihood estimation method without having to make any assumptions on the error of variances ratio, λ. Through simulation study, it is shown that the estimated values of the parameters are unbiased and consistent that indicated the adequacy of the proposed model. In addition, data examples also show the robustness aspects of the parameter estimates as it gives small values of standard deviation. In conclusion, the balanced replicated linear functional relationship model can be used in estimating the parameters where one can transform the unreplicated data to replicated data using general grouping or when replicated observations are available and their relationship can be described in a functional form. The novelty of the proposed model is one can overcome the unidentifiability problem in linear functional relationship model by grouping the observations to the unreplicated data and the assumption on the ratio of the error variances are no longer needed unlike in unreplicated linear functional relationship model.

CHAPTER 5: NONPARAMETRIC ESTIMATION FOR SLOPE OF BALANCED REPLICATED LINEAR FUNCTIONAL RELATIONSHIP MODEL

5.1 Introduction

In this chapter, a nonparametric method to estimate the slope parameter of balanced replicated linear functional relationship model is proposed where it addresses the third objective of the study. The motivation of this proposed method is to develop a robust estimator in estimating the slope parameter in the presence of outliers. The organization of the chapter is follows. Section 5.2 describes the general approach of nonparametric method in estimating a parameter while in Section 5.3 describes the propose new estimation method for a slope parameter of the balanced replicated linear functional relationship model using the nonparametric method. Description of simulation studies carried out to measure the performance of the proposed method is given in Section 5.4. This is followed by Section 5.5 that discusses the results of the simulation study. The proposed method is illustrated using real datasets namely Fat Mass Measurement data and Iron in Slag data in Section 5.6. Finally, summary and conclusion are given in Section 5.7.

5.2 Nonparametric Estimation Method of Linear Functional Relationship Model

Most of the methods in estimating the parameters of linear functional relationship model depends on the normality assumption (Fuller, 1987; Kendall & Stuart, 1979). In particular, for balanced replicated linear functional relationship model, the maximum likelihood estimation is a common method in estimating the parameters in which it has been shown on the need of a normality assumption as in Chapter 4. This means there will be problems if this assumption is not met. In other words, the parameter estimation of the slope parameter, for example, can lead to erroneous problems if the outliers present in the dataset. To overcome this problem, a robust method is deemed necessary and in this study, the nonparametric method is proposed to address the effect of outliers' presence. On that note, the interest of the study is on the balanced replicated linear functional relationship model where estimation of the slope parameter and thus is the scope of the study. This is because in most applications, the relationship between two linear variables can be described by the estimates of the slope.

Nonparametric estimation method is very popular due to its simplicity as it does not depend on a specific probability distribution. This method is widely used, easy to perform and also robust to outliers (Hajek, 1969). The nonparametric method is more efficient and it does depend on the normality assumption. Numerous studies have been carried out on the nonparametric estimation method; some examples include Dent (1935), Housner and Brennan (1948), Theil (1950) and Cheng and Ness (1999). However, not many consider the presence outlying observations in the datasets. From the literature review, Al-Nasser and Ebrahem (2005) and Ghapor et al. (2015) both have considered this situation where they incorporated the presence of outliers and have shown that the proposed methods are

robust to outliers, unlike all the other traditional methods. They also considered the varying the percentages of outliers present in the dataset from a single outlier up to 20% outliers. It is worthwhile to note that Al-Nasser and Ebrahem (2005) and Ghapor et al. (2015) methods are limited to unreplicated linear functional relationship model only. Therefore, it is worthwhile to explore if this approach can be improved and extended to accommodate a balanced replicated linear functional relationship model. In short, a new robust nonparametric method to estimate the slope parameter in balanced replicated linear functional relationship model is proposed by making some improvements to the nonparametric method as proposed by Al-Nasser and Ebrahem (2005) and Ghapor et al. (2015). The novelty of the proposed method is that the assumption of normality is not required and to estimate the slope parameter, the nonparametric approach is proposed. As usual, the proposed method is compared with the standard maximum likelihood estimation method.

5.3 A New Robust Nonparametric Estimation Method

In this section, the method of Al-Nasser and Ebrahem (2005) and Ghapor et al. (2015) is extended to balanced replicated linear functional relationship model. Instead of using median as in Al-Nasser and Ebrahem (2005) and Ghapor et al. (2015), the new proposed method uses the trimmed mean instead. The justification is that although the median is a robust indicator, particularly when the data has outliers, it only uses the 50^{th} percentiles of the observations and thus ignoring all values. Moreover, the trimmed mean is superior to median as it possesses a smaller asymptotic variance (Oosterhoff, 1994). Also, the standard error of the median is not very efficient as it contains a breakdown point of 50% (Hampel et al., 2011).

Trimmed mean is computed by averaging the values of the remaining data set after removing a certain percentage of the dataset (Wilcox, 2005). The unique feature of trimmed mean is that it calculates by discarding the lowest and highest $p\%$ of the values, then computing the mean of the remaining data. Based on the definition, trimmed mean is also considered as a robust estimates of location as a median. The trimmed mean is easy to compute and more efficient as it contains a small standard error.

Different authors used different percentage of trimming depending on their purposes of their research. For example, 10% trimming is used by Stigler (1973) and Ruppert and Carroll (1980) and they argued that the 10% trimming can be served as the best estimator. Hill and Dixon (1982) used several percentage of trimming namely 10%,15% and 20% trimming and found out that 15% is the best choice when the underlying distribution is unknown. The 20% trimming has been extensively investigated, and it frequently provides an appropriate balance between the mean and the median (Wilcox, 2005). In this study, the trimming used is 20% because numerous studies have shown that the effect of outliers present in the data will be diminished, thereby providing a reasonable estimate of the slope (Welsh, 1987; Wilcox, 2005). Hence, for the proposed method the trimmed mean, as opposed to the median, is used to calculate the slope parameter of the balanced replicated linear functional relationship model. As mentioned earlier, the proposed method is compared with the maximum likelihood method. The following are the description of the method proposed with six simple steps.

Firstly, the observed pairs (x_{ij}, y_{ij})'s, $i = 1,2, \dots, p\,; j = 1,2, \dots, m$ are ordered according to the magnitude of x value, assuming that all the x values are distinct. Next, sort these observations into several groups to obtain all the possible paired of slopes. The following step is to determine another possible paired of slopes by arranging the observed pairs according to the magnitude of y value. The steps involved are as follows:

Step 1: The observations are first arranged in ascending order, based on x value namely $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The associated values of y which may not be in ascending order are taken namely, $y_{[1]} \leq y_{[2]} \leq \cdots \leq y_{[n]}$ and obtain the new pairs $(x_{(i)}, y_{[j]})$.

Step 2: All the data are divided into p-subsamples that contains m elements such that $p \times m = n$. These subsamples can be rearranged in the form:

$$\begin{array}{cccc} (x_{(1)}, y_{[1]}) & (x_{(2)}, y_{[2]}) & \cdots & (x_{(m)}, y_{[m]}) \\ (x_{(m+1)}, y_{[m+1]}) & (x_{(m+2)}, y_{[m+2]}) & \cdots & (x_{(2m)}, y_{[2m]}) \\ \vdots & \vdots & \cdots & \vdots \\ (x_{(p-1)*(m+1)}, y_{[p-1\}*\{m+1]}) & \cdots & \cdots & (x_{(pm)}, y_{[pm]}) \end{array}$$

where p is the maximum divisor of n such that $p \leq m$. As an example, when $n = 40$, then $p = 5$ and $m = 8$. If the sample size is a prime number, then one can assume $p = 1$ and $m = n$ respectively.

Step 3: Find the number of all possible combination of paired slopes.

$$\left\{ b_x(k)_{ij} = \frac{y_{[j]} - y_{[i]}}{x_{(j)} - x_{(i)}};\ i = 1,2, \ldots, j-1; j = 2,3, \ldots, m \right\}; k = 1,2, \ldots, p$$

Step 4: Repeat Steps 1 to 3 by interchanging y and x to get possible paired of $b_y(k)_{ij}$

$$\left\{ b_y(k)_{ij} = \frac{y_{(j)} - y_{(i)}}{x_{[j]} - x_{[i]}};\ i = 1,2, \ldots, j-1; j = 2,3, \ldots, m \right\}; k = 1,2, \ldots, p$$

Step 5: Combine all the slopes from Step 4.

Step 6: Find the trimmed mean of the slopes $\hat{\beta}_{trim} = mean\{(b_x(k)_{ij}, b_y(k)_{ij}), trim = 20\%\}$.

Step 1 until Step 3 for estimating the slope parameter is based on the nonparametric estimation method as introduced by Al-Nasser and Ebrahem (2005) and Ghapor et al. (2015). Step 4 and Step 5 proposed by Ghapor et al. (2015) also incorporated in this method. However, in Step 6, the trimmed mean is used to calculate the slope parameter instead of median in Ghapor et al. (2015). This can be performed by removing the bottom and top 20% of the slopes and then calculating the mean of the remaining slopes values. This gives us a new slope parameter, $\hat{\beta}_{trim}$.

5.4 Simulation Studies

A simulation study is conducted in this section to compare the performance of the proposed nonparametric method with the maximum likelihood estimation method in the presence of outliers. First, simulated observations are obtained from the balanced replicated linear functional relationship model given by

$$Y = \alpha + \beta X_i, x_{ij} = X_i + \delta_{ij} \text{ and } y_{ij} = Y_i + \varepsilon_{ij} \quad (5.1)$$

where $X_i = 10\frac{i}{p}$ and the error terms $\delta_{ij}, \varepsilon_{ij} \sim N(0,0.1)$.

Without any loss of generality, the values of $\alpha = 1, \beta = 1$ and sample sizes, $n = 40, 80, 180,$ and 300 respectively are set. The cases comprised observations with no outlier, a single outlier, 5%, 10%, 15% and 20% outliers respectively. For each case, the simulation process was repeated 10000 times. The contaminated data points were generated as described by Al-Nasser and Ebrahem (2005) using this relationship

$$y_C = 1 + X_C + \varepsilon_C \text{ with } \varepsilon_C \sim N(0,25) \qquad (5.2)$$

Additionally, the robustness of the proposed method was also evaluated by generating the error terms from three different cases which included the non-normal symmetric case, the right-skewed case, and the left-skewed case. These cases were generated from the beta distribution with parameters (3,3), (2,9) and (9,2) respectively using the same relationship described earlier. The probability density function of the beta distribution as follows:

$$f(x) = \frac{1}{Beta(a,b)} x^{(a-1)}(1-x)^{(b-1)}, \qquad 0 \leq x \leq 1$$

where a and b are two positive shape parameters, and $Beta(a,b)$ can be defined as the beta function.

The performance of both methods was examined by observing the estimated bias (EB) and the mean square error (MSE) of the slope parameter. The estimated bias and the mean square error are defined by

$$\text{Estimated Bias, EB} = |\widehat{w} - w| \text{ and} \tag{5.3}$$

$$\text{Mean Square Error, MSE} = \frac{1}{s}\sum\left(\widehat{w}_j - w\right)^2 \tag{5.4}$$

where w be a generic term for the parameters. Table 5.1 represents the required values of p and m for the proposed estimator for this model.

Table 5.1 Values of p and m

Sample size, n	Subgroups, p	Number of elements, m
40	5	8
80	8	10
180	12	15
300	15	20

5.5 Results and Discussion

Results of the simulation studies are presented where performance based on the estimated bias and the mean square error are summarized in Table 5.2 to Table 5.5 and Table 5.6 to Table 5.9 respectively. Table 5.2 to Table 5.5 show the estimated bias for the slope of the replicated linear functional relationship model when the errors δ_{ij} and ε_{ij} are from normal distribution, beta distribution with parameter (2,9), beta distribution with parameter (9,2) and beta distribution with parameter (3,3).

Table 5.2 Estimated Bias of the Slope: Normal Case: Normal (0,0.1)

Contamination	Method	Sample Size			
		40	80	180	300
No outlier	MLE	3.088E-04	2.570E-04	1.717E-05	1.012E-05
	Proposed	1.182E-02	2.602E-03	2.868E-03	4.463E-03
Single outlier	MLE	2.928E-01	1.484E-01	5.940E-02	2.935E-02
	Proposed	2.592E-02	9.250E-03	6.006E-03	6.400E-03
5%	MLE	4.496E-01	3.633E-01	1.725E-01	1.584E-02
	Proposed	3.641E-02	2.081E-02	1.143E-02	3.468E-03
10%	MLE	2.846E-01	1.714E-02	2.355E-01	1.579E-02
	Proposed	3.070E-02	1.353E-03	1.594E-02	1.941E-03
15%	MLE	3.412E+00	6.060E+00	1.726E-01	1.578E-02
	Proposed	4.427E-02	2.700E-02	1.306E-02	8.107E-04
20%	MLE	2.025E+01	1.712E-02	1.607E-02	1.577E-02
	Proposed	1.539E-01	7.564E-03	9.270E-03	6.706E-03

From Table 5.2 for normal case, the estimated bias for the slope shows the maximum likelihood estimation method is superior when the data has no outliers. However, when a single outlier to 20% outliers are introduced in the data set, the maximum likelihood estimation tends to break down quickly and become huge. The proposed method however, consistently shows a small bias measure with the increasing percentage of outliers as well as the increasing of sample sizes.

Table 5.3 Estimated Bias of the Slope: Right Skewed Case: Beta (2,9)

Contamination	Method	Sample Size			
		40	80	180	300
No outlier	MLE	1.766E-05	4.770E-05	5.919E-05	1.764E-06
	Proposed	4.997E-03	1.827E-03	9.317E-04	8.804E-04
Single outlier	MLE	2.918E-01	1.482E-01	5.911E-02	2.918E-02
	Proposed	1.021E-02	4.208E-03	2.019E-03	1.536E-03
5%	MLE	4.481E-01	3.625E-01	1.713E-01	1.674E-02
	Proposed	1.429E-02	8.576E-03	4.291E-03	7.501E-04
10%	MLE	2.919E-01	1.837E-02	2.343E-01	1.682E-02
	Proposed	1.292E-02	1.353E-03	6.127E-03	6.739E-04
15%	MLE	1.054E+00	7.693E+00	1.714E-01	1.699E-02
	Proposed	2.741E-02	1.246E-02	5.809E-03	4.324E-04
20%	MLE	2.000E+01	1.836E-02	1.744E-02	1.700E-02
	Proposed	1.444E-01	2.046E-03	1.765E-04	9.158E-04

Table 5.4 Estimated Bias of the Slope: Left Skewed Case: Beta (9,2)

Contamination	Method	Sample Size			
		40	80	180	300
No outlier	MLE	6.406E-05	5.297E-05	5.010E-05	3.911E-05
	Proposed	4.970E-03	1.745E-03	1.058E-03	8.541E-04
Single outlier	MLE	2.890E-01	1.454E-01	5.666E-02	2.635E-02
	Proposed	1.023E-02	4.149E-03	2.154E-03	1.507E-03
5%	MLE	4.453E-01	3.598E-01	1.687E-01	1.989E-02
	Proposed	1.425E-02	8.510E-03	4.443E-03	7.020E-04
10%	MLE	2.890E-01	2.117E-02	2.317E-01	1.971E-02
	Proposed	1.293E-02	1.233E-03	6.287E-03	6.297E-04
15%	MLE	4.801E+00	1.034E+01	1.688E-01	1.971E-02
	Proposed	2.758E-02	1.225E-02	5.922E-03	3.961E-04
20%	MLE	2.040E+01	2.118E-02	2.001E-02	1.972E-02
	Proposed	1.452E-01	1.858E-03	2.928E-04	9.295E-04

Table 5.5 Estimated Bias of the Slope: Non-normal Symmetric Case: Beta (3,3)

Contamination	Method	Sample Sizes			
		40	80	180	300
No outlier	MLE	9.630E-05	2.467E-04	7.845E-06	4.652E-05
	Proposed	7.877E-03	2.186E-03	1.084E-03	1.004E-03
Single outlier	MLE	2.903E-01	1.464E-01	5.803E-02	2.779E-02
	Proposed	1.671E-02	6.307E-03	2.951E-03	2.168E-03
5%	MLE	4.467E-01	3.610E-01	1.702E-01	1.820E-02
	Proposed	2.336E-02	1.370E-02	6.668E-03	5.133E-04
10%	MLE	2.906E-01	1.983E-02	2.331E-01	1.818E-02
	Proposed	2.044E-02	7.474E-04	9.461E-03	8.216E-05
15%	MLE	3.825E+00	8.835E+00	1.702E-01	1.818E-02
	Proposed	3.468E-02	1.890E-02	8.027E-03	1.077E-03
20%	MLE	2.031E+01	1.982E-02	1.857E-02	1.816E-02
	Proposed	1.494E-01	1.475E-04	3.116E-03	4.973E-03

Similar results can be found in Table 5.3, Table 5.4 and Table 5.5. In general, from all four cases, as the sample size increase from 40 to 300, the estimated bias is decreasing for both maximum likelihood estimation method and the proposed nonparametric method. The same can be with the introduction of outliers in the data from single outlier to 20% outliers, the estimated bias decreases as the number of observations increase. On the other hand, the proposed method gives smaller estimated bias value compared to the maximum likelihood estimation method when the observations have outliers.

Table 5.6 Mean Square Error of the Slope: Normal Case: Normal (0,0.1)

Contamination	Method	Sample Size			
		40	80	180	300
No outlier	MLE	6.348E-04	3.058E-04	1.365E-04	8.018E-05
	Proposed	8.087E-04	3.252E-04	1.482E-04	1.017E-04
Single outlier	MLE	8.623E-02	2.229E-02	3.661E-03	9.416E-04
	Proposed	1.362E-03	4.092E-04	1.768E-04	1.230E-04
5%	MLE	2.026E-01	1.323E-01	2.987E-02	3.344E-04
	Proposed	2.032E-03	7.589E-04	2.736E-04	9.850E-05
10%	MLE	1.410E-01	6.100E-04	5.558E-02	3.325E-04
	Proposed	1.711E-03	3.749E-04	4.064E-04	9.443E-05
15%	MLE	3.744E+01	6.439E+01	2.991E-02	3.323E-04
	Proposed	3.240E-03	1.099E-03	3.355E-04	9.974E-05
20%	MLE	4.230E+02	6.094E-04	3.997E-04	3.321E-04
	Proposed	2.942E-02	5.125E-04	2.808E-04	1.616E-04

From Table 5.6 where the error of variances is normally distributed, the mean square error (MSE) for maximum likelihood estimation method gives better result than the proposed method for each sample size when the data has no outlier. However, when a single outlier to 20% outliers are introduced in the data, the proposed method shows consistently smaller values of mean square error compared to the maximum likelihood estimation method for each sample size.

Table 5.7 Mean Square Error of the Slope: Right Skewed Case: Beta (2,9)

Contamination	Method	Sample Size			
		40	80	180	300
No outlier	MLE	7.748E-05	3.833E-05	1.699E-05	1.011E-05
	Proposed	1.071E-04	4.225E-05	1.777E-05	1.078E-05
Single outlier	MLE	8.521E-02	2.199E-02	3.512E-03	8.623E-04
	Proposed	1.922E-04	5.761E-05	2.120E-05	1.245E-05
5%	MLE	2.008E-01	1.314E-01	2.938E-02	2.903E-04
	Proposed	2.973E-04	1.149E-04	3.603E-05	1.119E-05
10%	MLE	8.525E-02	3.777E-04	5.492E-02	2.941E-04
	Proposed	2.702E-04	4.688E-05	5.649E-05	1.158E-05
15%	MLE	9.002E+00	7.007E+01	2.938E-02	2.995E-04
	Proposed	1.229E-03	2.030E-04	5.391E-05	1.207E-05
20%	MLE	4.019E+02	3.775E-04	3.221E-04	2.999E-04
	Proposed	2.585E-02	5.604E-05	2.161E-05	1.351E-05

Looking at the Table 5.7, where the error variances, δ_{ij} and ε_{ij} are skewed to the right with beta distribution (2,9), again the maximum likelihood estimation method performs well when no outlier exist in the data. However, the maximum likelihood estimation method breaks down when a single outlier until 20% outliers are present in the data set. The proposed method gives better result in estimating the slope parameter.

Table 5.8 Mean Square Error of the Slope: Left Skewed Case: Beta (9,2)

Contamination	Method	Sample Size			
		40	80	180	300
No outlier	MLE	7.741E-05	3.817E-05	1.675E-05	1.019E-05
	Proposed	1.072E-04	4.202E-05	1.781E-05	1.082E-05
Single outlier	MLE	8.357E-02	2.118E-02	3.226E-03	7.041E-04
	Proposed	1.917E-04	5.718E-05	2.159E-05	1.241E-05
5%	MLE	1.983E-01	1.295E-01	2.847E-02	4.059E-04
	Proposed	2.942E-04	1.136E-04	3.726E-05	1.122E-05
10%	MLE	8.360E-02	4.886E-04	5.371E-02	4.001E-04
	Proposed	2.674E-04	4.773E-05	5.836E-05	1.171E-05
15%	MLE	4.111E+01	1.077E+02	2.850E-02	3.995E-04
	Proposed	1.235E-03	1.987E-04	5.549E-05	1.207E-05
20%	MLE	4.179E+02	4.889E-04	4.185E-04	3.998E-04
	Proposed	2.608E-02	5.745E-05	2.164E-05	1.377E-05

Table 5.9 Mean Square Error of the Slope: Non-normal Symmetric Case: Beta (3,3)

Contamination	Method	Sample Size			
		40	80	180	300
No outlier	MLE	2.252E-04	1.083E-04	4.771E-05	2.877E-05
	Proposed	3.106E-04	1.216E-04	5.167E-05	3.116E-05
Single outlier	MLE	8.448E-02	2.153E-02	3.415E-03	8.012E-04
	Proposed	5.403E-04	1.594E-04	5.983E-05	3.513E-05
5%	MLE	1.997E-01	1.304E-01	2.900E-02	3.607E-04
	Proposed	8.147E-04	3.094E-04	9.657E-05	3.206E-05
10%	MLE	8.461E-02	5.068E-04	5.438E-02	3.606E-04
	Proposed	7.091E-04	1.367E-04	1.446E-04	3.312E-05
15%	MLE	3.641E+01	8.952E+01	2.902E-02	3.609E-04
	Proposed	1.907E-03	4.931E-04	1.229E-04	3.630E-05
20%	MLE	4.170E+02	5.060E-04	3.949E-04	3.602E-04
	Proposed	2.749E-02	1.540E-04	7.304E-05	6.305E-05

The same can be said when looking at Table 5.8 where the error variances, δ_{ij} and ε_{ij} are skewed to the left with beta distribution (9,2) and Table 5.9 where the error variances, δ_{ij} and ε_{ij} are non-normal with beta distribution (3,3). Thus, it can be concluded that the nonparametric method is superior than the maximum likelihood estimation method for estimating the slope parameter in the presence of outliers.

Based on the simulation results, it is clearly shows the advantages of using the proposed robust nonparametric method namely the 20% trimmed mean in estimating the slope parameter. The value of the estimated bias and the mean square errors of the 20% trimmed mean are generally less than the maximum likelihood estimation method when

the datasets have a single outlier and a certain percentage of outliers. By comparing with the maximum likelihood estimation method in which it is sensitive to the outliers, the 20% trimmed mean is the best estimator as it is the compromise between the mean and median under both symmetric and non-symmetric distribution.

5.6 Examples

In this section, the proposed nonparametric method is applied to published dataset and compared with the traditional maximum likelihood estimation method in estimating the slope parameter. Two data sets are used namely the Fat Mass Measurement data and Frosted Flakes data. Measurement error are assumed to occur in both variables to make the relationship as given in (4.1).

5.6.1 Fat Mass Measurements Data

By considering a data set from Goran et. al (1996), the data set consists of 96 observations that are free from outliers. As measurement error can occur in both variables for this experiment, it is noted the relationship between two variables can be described by balanced replicated linear functional relationship model as given in (4.1). Here, it is assumed that the error terms follow a normal distribution. Since there are 96 observations in this data, the group of the data is obtained by dividing the data into 8 groups and each groups have 12 observations that are balanced and equal in each group in order to estimate the slope parameter by the proposed method. The data can be modelled by

$$Y_i = \alpha + \beta X_i$$

for $i = 1,2,\cdots,8$ and $j = 1,2,\cdots,12$ and $n = p \times m = 8 \times 12 = 96$. To create different situations in investigating the slope effect by two different methods, some original y values are substituted by outliers namely a single outlier, 5%, 10% and 15% outliers by following Kim (2000) and Imon & Hadi (2008). The estimated slopes (and standard deviation) by using two different methods are shown in Table 5.10.

Table 5.10 Slopes Estimates Using Fat Mass Measurements Data

Contamination	MLE, $\hat{\beta}_{MLE}$ (Standard Deviation)	Proposed Method, $\hat{\beta}_{trim}$ (Standard Deviation)
No outlier	1.097 (0.512)	0.974 (0.452)
Single outlier	1.483 (0.993)	1.001 (0.603)
5% outliers	4.041 (6.660)	1.014 (0.861)
10% outliers	6.561(25.790)	1.023 (1.013)
15% outliers	9.527(26.872)	1.002 (1.052)

From Table 5.10, both methods showed a somewhat similar value of the slope estimates which approximately equal to one when the dataset has no outlier. However, when outliers increased from a single outlier to 15%, the estimates of the slope using the maximum likelihood method becomes huge compared to the proposed nonparametric method. By comparing the standard deviation for both methods, the slope estimate using the proposed method has a smaller standard deviation when the data has no outlier up to

15% outliers. This clearly shows the proposed method works well in estimating the slope parameter when the dataset contains outliers.

5.6.2 Iron in Slag Data

To demonstrate the practicality of the proposed method, another dataset called the iron in slag data was utilized. Details on the dataset is given in APPENDIX D. The data set consists of 50 observations that are free from outliers. As measurement error can occur in both variables for this experiment, it is assumed that the error terms follow a normal distribution and the relationship can be described by replicated linear functional relationship model as given in (4.1). To apply the proposed method, the data set are divided into 5 groups with 10 observations in each group. The data can be modelled by

$$Y_i = \alpha + \beta X_i$$

for $i = 1,2,\cdots,5$ and $j = 1,2,\cdots,10$ and $n = p \times m = 5 \times 10 = 50$. To create different conditions in investigating the slope effect by two different methods, some original y values are substituted by outliers namely a single outlier, 5%, 10% and 15% outliers by following Kim (2000) and Imon & Hadi (2008). The estimated slopes (and standard deviation) by using two different methods are shown in Table 5.11.

Table 5.11 Slope estimates using the iron in slag dataset

Contamination	MLE, $\hat{\beta}_{MLE}$ (Standard Deviation)	Proposed Method, $\hat{\beta}_{trim}$ (Standard Deviation)
No outlier	0.922 (0.947)	0.835 (0.864)
Single outlier	1.223 (1.513)	0.876 (1.108)
5% outliers	2.435 (2.932)	0.901 (1.214)
10% outliers	5.449 (16.293)	1.038 (1.707)
15% outliers	9.534 (29.428)	1.268 (2.458)

As shown in Table 5.11, in the absence of outliers, the slope estimates obtained using the proposed method are almost identical to the maximum likelihood estimation method. However, in the presence of a single outlier, 5%, 10%, and 15% outliers, the slope estimates for maximum likelihood estimation method are shown to increase, although the proposed estimator is not considerably affected by the outliers. The standard deviations for the slope estimate using the proposed method are smaller compared to the slope estimates using the maximum likelihood estimation method. This clearly demonstrates that the proposed method performs well in estimating the slope parameter when the dataset contains outliers.

5.7 Summary and Conclusions

In this chapter, a 20% trimmed mean as the robust nonparametric method in estimating the slope parameter of balanced replicated linear functional relationship model is proposed. By looking at the estimated bias and the mean square, it is concluded that

the proposed nonparametric method is superior to the maximum likelihood estimation in the presence of outliers. This can be seen from the simulation studies when the percentage of outliers increases, the estimated bias and the mean square error of the maximum likelihood estimation becomes huge and breaks down easily as compared to the proposed nonparametric method. The estimated bias and the mean square error of the 20% trimmed mean are not affected by outliers regardless of the percentage of the contamination or the sample sizes. The same results also can be seen in real data examples.

Although the maximum likelihood estimation method is a common method used in estimating the slope of a balanced replicated linear functional relationship model, the assumption of normality in data sets that contain outliers leads to errors in estimating the parameters of the model particularly the slope parameter. In summary, the new nonparametric approach proposed by using the 20% trimmed mean can be viewed as the robust estimator in estimating the slope parameters in balanced replicated linear functional relationship model even when the datasets have very large percentages of outliers. This method also can be utilized as an alternative method to estimate the slope parameter of the balanced replicated linear functional relationship model in the presence of outliers. The novelty of this nonparametric method is that it does not require any assumption on the probability distribution of the data and also is easy to apply. Unlike median, the trimmed mean takes into account of the 80% of the observations.

CHAPTER 6: SINGLE OUTLIER DETECTION FOR BALANCED REPLICATED LINEAR FUNCTIONAL RELATIONSHIP MODEL

6.1 Introduction

This chapter addresses the fourth objective of the study that is the outlier detection in a balanced replicated linear functional relationship model. Identifying potential outliers is important as their presence leads to skewed or bias results. The motivation of this study is to provide a technique that can be used in a data quality evaluation process for data that fits linear functional relationship model. Outlier detection procedure for balanced replicated linear functional relationship model has not been explored before. Here, *COVRATIO* statistics as an outlier detection method is considered in which this method is modified to accommodate for a balanced replicated linear functional relationship model. In Section 6.2, the formulation of *COVRATIO* statistic for balanced replicated linear functional relationship model is described. Next, the procedure for determining the cut-off points for detecting an outlier is given in Section 6.3. In Section 6.4, the power of performance is discussed. Section 6.5 illustrates some practical applications of the *COVRATIO* statistic using simulated and real data. Finally, summary and conclusion are given in Section 6.6.

6.2 *COVRATIO* Statistic for Balanced Replicated Linear Functional Relationship Model

Outliers are any observations that do not follow any pattern in the dataset. The existence of outliers may affect the parameter estimation of the models and also the prediction of the analysis (Satari & Khalif, 2020). Identification of outliers becomes a vital area in which many outlier detection methods have been proposed including the *COVRATIO* statistic. *COVRATIO* statistic has been used for different models such as in a linear and circular regression model (Belsley et al., 1980; Abuzaid et al., 2011; Ibrahim et al., 2013). Moreover, in errors-in-variable model, the *COVRATIO* statistic has been utilized in a structural relationship model and also in a functional relationship model (Hussin et al., 2010; Ghapor et al., 2014; Mamun et al., 2019; Mokhtar et al., 2019).

The *COVRATIO* statistic has been first introduced by Belsley et al. (1980) in identifying influential observations or outliers in a linear regression model. The idea of the *COVRATIO* statistic is based on the determinantal ratio of covariance matrix for a full data set and a reduced data set by excluding one observation in turn. In other words, i^{th} is deleted from the full data set. It is given by following:

$$\left|COVRATIO_{(-i)} - 1\right| = \frac{|COV|}{\left|COV_{(-i)}\right|} \tag{6.1}$$

where $|COV|$ is the determinant of covariance matrix for full data set and $\left|COV_{(-i)}\right|$ is for determinant of covariance matrix for the reduced data set by excluding the i^{th} row. If the ratio is close to 1, then there is no significant difference between the covariance matrices.

Otherwise, the i^{th} observation is consistent with other observations. Moreover, they also developed a test statistic of the form $\left|COVRATIO_{(-i)} - 1\right|$ and identified the cut-off point for examining the presence of the outliers.

For balanced replicated linear functional relationship model, the ratio of statistic is based on the determinant of the asymptotic variance and covariance for the parameters that has been discussed in Chapter 4 (4.31). It is motivated by the fact that the balanced replicated linear functional relationship model has a closed form covariance matrix of the parameters. Algebraically the *COVRATIO* statistic for balanced replicated linear functional relationship model is given by

$$\left|COVRATIO_{(-i)} - 1\right| = \frac{|COV|}{\left|COV^{*}_{(-i)}\right|} \tag{6.2}$$

where $|COV|$ is the determinant of covariance matrix for full data set and $\left|COV^{*}_{(-i)}\right|$ is the determinant of covariance matrix by deleting i^{th} observation of every group.

The deleted i^{th} observation is replaced with mean samples to make balanced replication of all sample groups. The use of mean substitution may be based on the fact that the mean is a reasonable guess of a value for a randomly selected observation from a normal distribution (Acock, 2005). Any observation with $\left|COVRATIO_{(-i)} - 1\right|$ exceeds the cut-off points will be considered as an outlier. The cut-off points are obtained through the simulation studies in the following section.

6.3 Determination of Cut-off Points by *COVRATIO* Statistic

A simulation study is conducted to obtain the cut-off points of *COVRATIO* statistic for balanced replicated linear functional relationship model. Eight different sample sizes $n = 20, 40, 60, 80, 100, 132, 180$ and 300 are used according to the division of sample size as in Table 6.1.

Table 6.1 Values of groups and elements

Sample size, n	Subgroups, p	Number of elements, m
20	4	5
40	5	8
60	6	10
80	8	10
100	10	10
132	11	12
180	12	15
300	15	20

Furthermore, different values of $\tau^2 = 0.2, 0.4, 0.6, 0.8$ and 1.0 are chosen. For each sample of size n and τ^2, a set of normal random errors are generated from the normal distribution with mean 0 and τ^2 respectively. Thus, to obtain the cut-off points of *COVRATIO* statistic, the following steps are proposed:

Step 1: Generate a fixed variable $X_i = 10(i/p)$ of size p, with $i = 1,2, \dots, p$ where p is the number of group. Without loss of generality, the intercept, slope and error variance parameters of replicated linear functional relationship model are fixed at $\alpha = 0$, $\beta = 1$ and $\sigma^2 = 1$ respectively.

Step 2: Generate two random error terms δ_{ij} and ε_{ij} from $N(0, \sigma^2)$ and $N(0, \tau^2)$ respectively.

Step 3: Calculate the observed values of x_{ij} and y_{ij} using equation (4.1).

Step 4: Fit the generated data to balanced replicated linear functional relationship model and estimate the parameters of balanced replicated linear functional relationship model.

Step 5: Find the variance-covariance matrix and calculate the $|COV|$ for all data.

Step 6: Delete the i^{th} observation of every group and replicate with mean for observation in every group from the generated sample of both x_{ij} and y_{ij} where $i = 1,2, \dots, p$ and $j = 1,2, \dots, m$. Repeat steps 4 till steps 6 to obtain $\left|COV_{(-i)}\right|$.

Step 7: Calculate $COVRATIO_{(-i)}$ and find the value of $\left|COVRATIO_{(-i)} - 1\right|$ for all i.

Step 8: Specify the maximum value of $\left|COVRATIO_{(-i)} - 1\right|$.

The procedures are simulated 5000 times for each combination of sample size n and τ^2. Then, the 1%, 5% and 10% upper percentiles of the maximum values of $\left|COVRATIO_{(-i)} - 1\right|$ are calculated. These upper percentiles are used as the cut-off points in identifying the outliers for the balanced replicated linear functional relationship model.

The cut-off points for various sample sizes of n are tabulated at 1%, 5 % and 10% levels of significant as given in Table 6.2, Table 6.3 and Table 6.4 respectively. The 1%,

5% and 10% upper percentile values of $\left|COVRATIO_{(-i)} - 1\right|$ are independent of τ^2 for all n. From all the tables, the cut-off points show a decreasing function of sample size n.

Table 6.2 The 1% upper percentile points of $\left|COVRATIO_{(-i)} - 1\right|$

Sample size, n	$\tau^2 = 0.2$	$\tau^2 = 0.4$	$\tau^2 = 0.6$	$\tau^2 = 0.8$	$\tau^2 = 1.0$
20	2.9727	4.2716	5.1172	6.0847	6.4157
40	2.1920	2.1136	1.9942	2.0638	2.1304
60	0.9880	1.0325	1.0345	1.0792	1.1230
80	0.9710	0.9312	0.9060	0.8770	0.8567
100	0.9493	0.9008	0.8542	0.8213	0.7930
132	0.9111	0.8256	0.7724	0.7280	0.6974
180	0.8410	0.7335	0.6624	0.6189	0.5826
300	0.6893	0.5563	0.4876	0.4448	0.4116

Table 6.3 The 5% upper percentile points of $\left|COVRATIO_{(-i)} - 1\right|$

Sample size, n	$\tau^2 = 0.2$	$\tau^2 = 0.4$	$\tau^2 = 0.6$	$\tau^2 = 0.8$	$\tau^2 = 1.0$
20	1.1321	1.9974	2.5915	2.8700	3.0111
40	1.3591	1.2633	1.3131	1.3179	1.3985
60	0.9804	0.9529	0.9281	0.9078	0.8913
80	0.9625	0.9158	0.8807	0.8498	0.8227
100	0.9401	0.8797	0.8322	0.7931	0.7632
132	0.8967	0.8086	0.7491	0.7049	0.6705
180	0.8245	0.7139	0.6431	0.5977	0.5613
300	0.6741	0.5412	0.4719	0.4288	0.3969

Table 6.4 The 10% upper percentile points of $\left|COVRATIO_{(-i)} - 1\right|$

Sample size, n	$\tau^2 = 0.2$	$\tau^2 = 0.4$	$\tau^2 = 0.6$	$\tau^2 = 0.8$	$\tau^2 = 1.0$
20	0.9998	1.3477	1.7647	1.9923	2.1651
40	1.0873	1.0473	1.0578	1.0783	1.1078
60	0.9770	0.9453	0.9171	0.8924	0.8717
80	0.9581	0.9084	0.8690	0.8372	0.8083
100	0.9347	0.8700	0.8195	0.7797	0.7492
132	0.8898	0.7987	0.7369	0.6915	0.6565
180	0.8173	0.7026	0.6309	0.5855	0.5501
300	0.6668	0.5330	0.4636	0.4199	0.3891

In order to obtain the cut-off points for balanced replicated linear functional relationship model, the arithmetic mean of the values $\left|COVRATIO_{(-i)} - 1\right|$ for each τ^2 at 1%, 5% and 10% significance level is calculated. Then, the curve is plotted as shown in Figure 6.1, Figure 6.2 and Figure 6.3 respectively. The equation of the series trend line is obtained by fitting the curve with the power series equation; for example, in Figure 6.2, at 5% significance level, the equation for cut-off point is $y = 9.6293n^{-0.526}$ where n is the sample size. Similar formulations of the trend lines are obtained for 1% and 10% significant level as in Figure 6.1 and Figure 6.3. At 5% and 10% significance level, the curve have a good fit of R^2 which is approximately equal to 1. At 1% significance level, the R^2 shows approximately equal to 0.9 which is still considered a good fit.

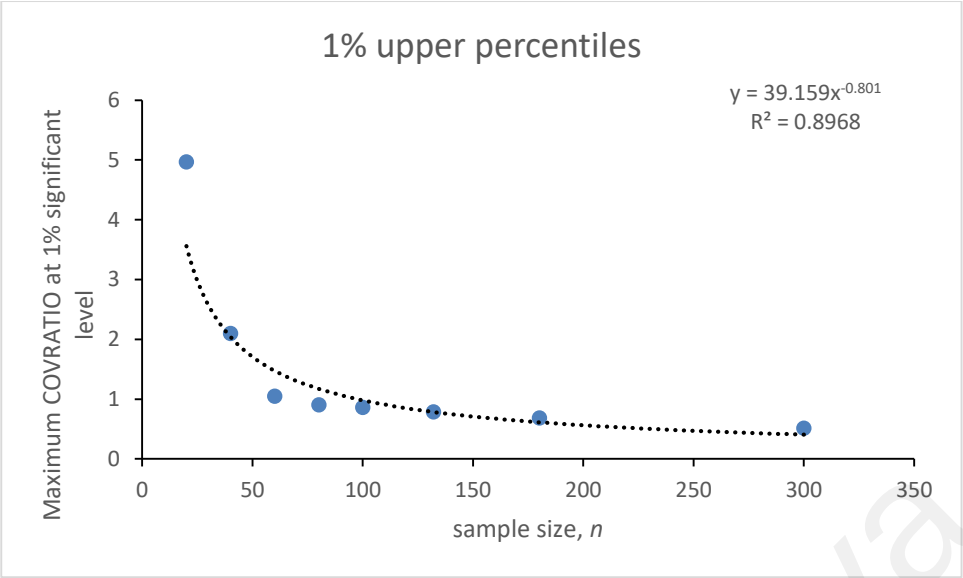

Figure 6.1 Graph of the Power Series in Finding the General Formula for the Cut-off Point at 1% Significant Level

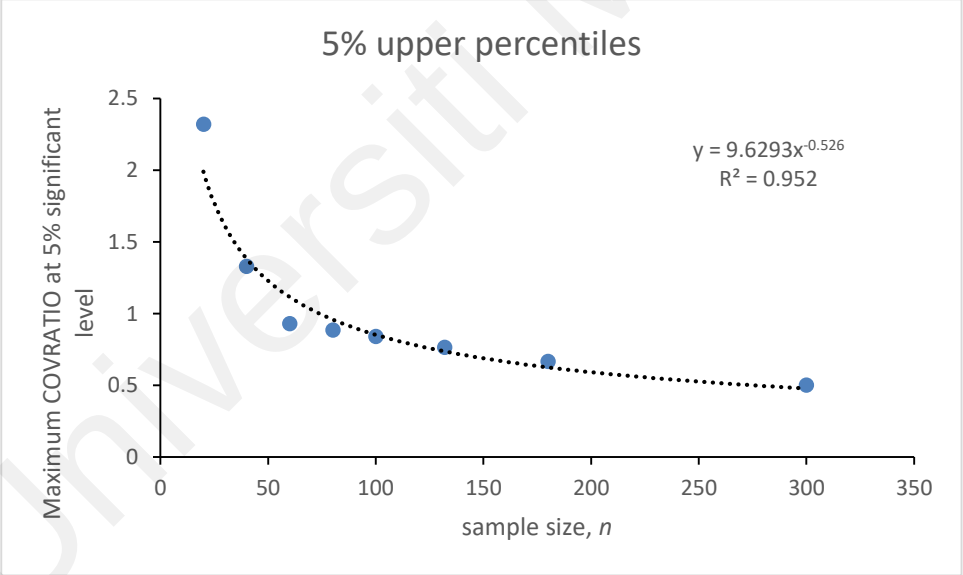

Figure 6.2 Graph of the Power Series in Finding the General Formula for the Cut-off Point at 5% Significant Level

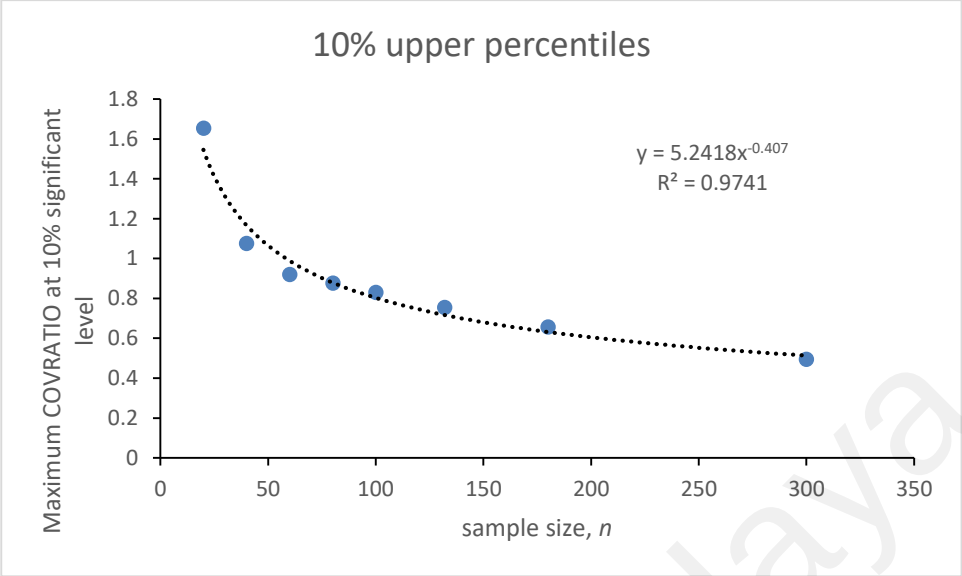

Figure 6.3 Graph of the Power Series in Finding the General Formula for the Cut-off Point at 10% Significant Level

Table 6.5 General formula for cut-off points at 1%, 5% and 10% upper percentile, where n is the sample size

Upper Percentile	General Formula for Cut-off Points
10%	$y = 5.2418n^{-0.407}$
5%	$y = 9.6293n^{-0.526}$
1%	$y = 39.159n^{-0.801}$

Table 6.5 shows the general formula for cut-off point in detecting the outliers at 1%, 5% and 10% upper percentile where n is the sample size. Based on these formulas, any point that exceeds the cut-off points will be considered as an outlier.

6.4 Power of Performance for *COVRATIO* Statistic

The next step of the analysis is to measure the power of performance of $\left|COVRATIO_{(-i)} - 1\right|$ using Monte Carlo simulation method. Different sample of sizes $n = 40, 80, 100$ and 180 according to division of sample size as given in Table 6.1 in sub-chapter 6.3 are considered for this study. The procedure to generate data set as described in Section 6.3 is applied here. In addition, the contamination is randomly applied at randomly chosen an observation, for example at position c. Then, y_C is contaminated as follows:

$$y_C = \alpha + \beta X_C + \varphi_C, \tag{6.3}$$

where y_C and X_C are the value of the c^{th} position of both variables y and X respectively after contamination. In addition, φ_C is error taken from normal distribution with mean zero and different variances of 6,8,10,12,14 and 16 respectively. The data generated is fitted by using the model in (4.1) and then $|COV|$ is calculated. Then, the i^{th} observation is deleted for every group and replaced with the mean of the remaining data. This to ensure the elements in each group still balanced and under replicated model. The data is refitted and $COVRATIO_{(-i)}$ is calculated. The maximum value of $\left|COVRATIO_{(-i)} - 1\right|$ is specified and compared with the specified cut-off point. As mentioned in the previous section, the 5% significance level as the cut-off point is used to detect the presence of the outlier with 95% confidence interval. This procedure has correctly identified the outlier in the data set if the values of $\left|COVRATIO_{(-i)} - 1\right|$ is maximum and exceeds the stated cut-off point. The process is repeated 5000 times. The power of performance is then

examined by calculating the percentage of the correct detection of the contaminated observation at c^{th} position.

Figure 6.4 to Figure 6.7 show the graph of power of performance of $\left|COVRATIO_{(-i)} - 1\right|$ statistic for $n = 40,80,100$ and 180 respectively with different level of $\tau^2 = 0.2,0.4,0.6,0.8$ and 1.0. From these plots, it can be concluded that as τ^2 decreases, the power of performance in detecting the correct outlier increases for all n.

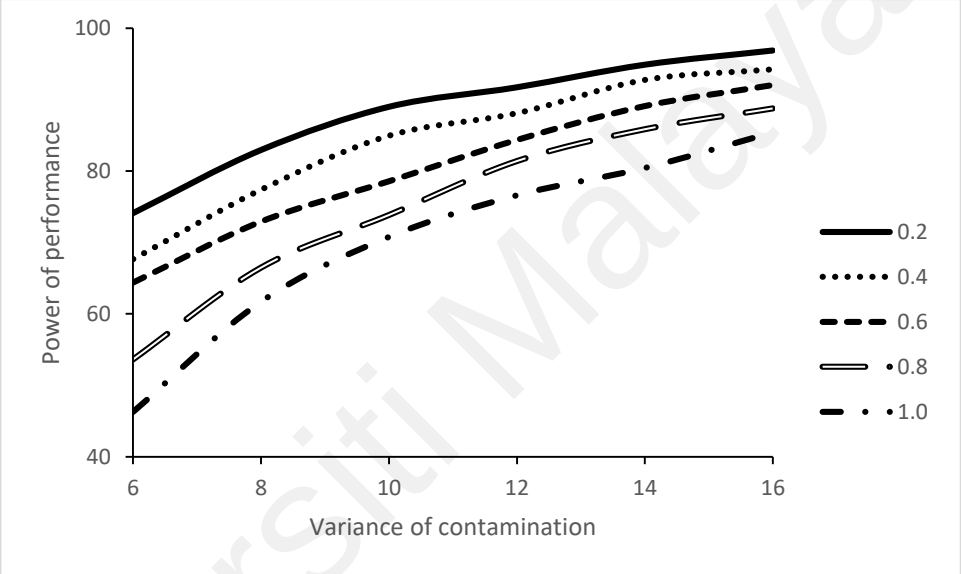

Figure 6.4 Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $n = 40$

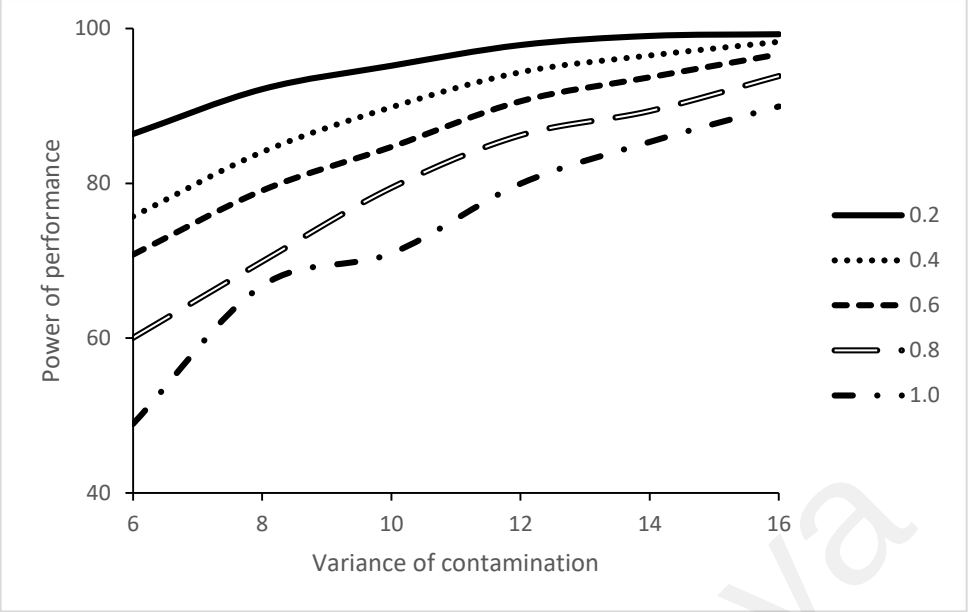

Figure 6.5 Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $n = 80$

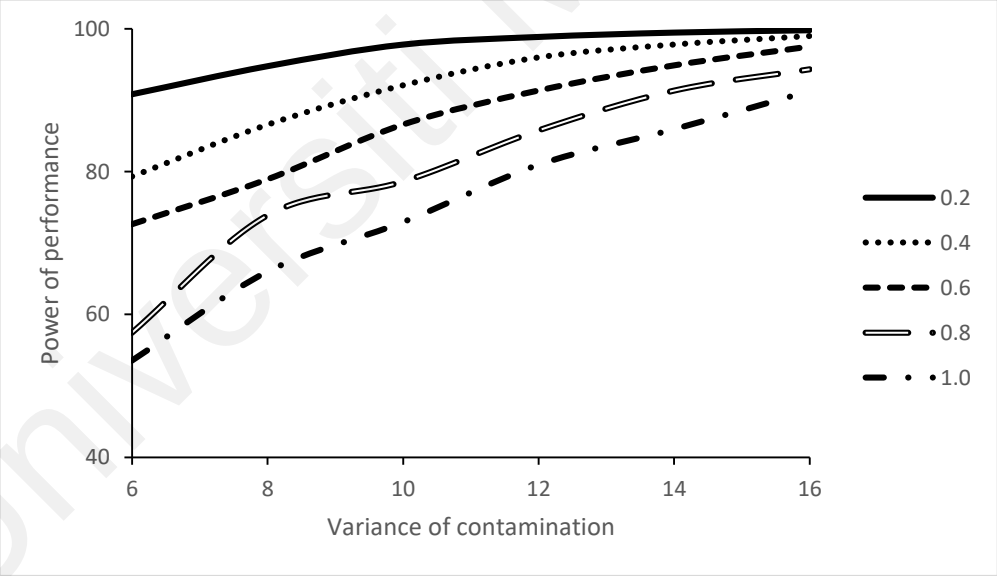

Figure 6.6 Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $n = 100$

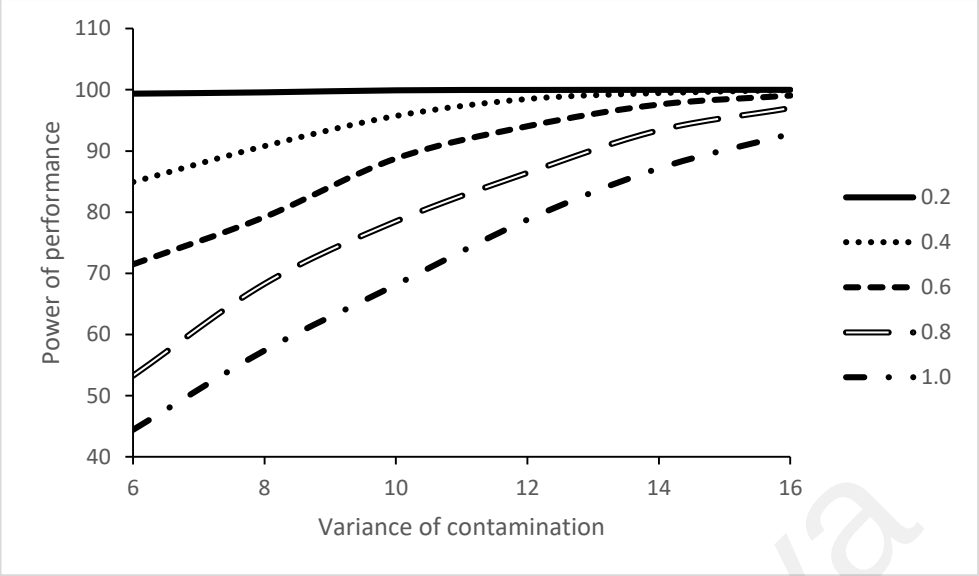

Figure 6.7 Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $n = 180$

Also Figure 6.8 shows the power of performance of $\left|COVRATIO_{(-i)} - 1\right|$ statistic for $\tau^2 = 0.2$ for different sample sizes namely when $n = 40,80,100$ and 180. From this figure, the power of performance increases as the variance of contamination increases. By looking at this figure, one can see that the power of performance is independent of sample size. Similar trends can be obtained in Figure 6.9 to Figure 6.12 when $\tau^2 = 0.4,0.6,0.8$ and 1.0 respectively. These figures also give consistent results whereby the power of performance is independent for all sample sizes.

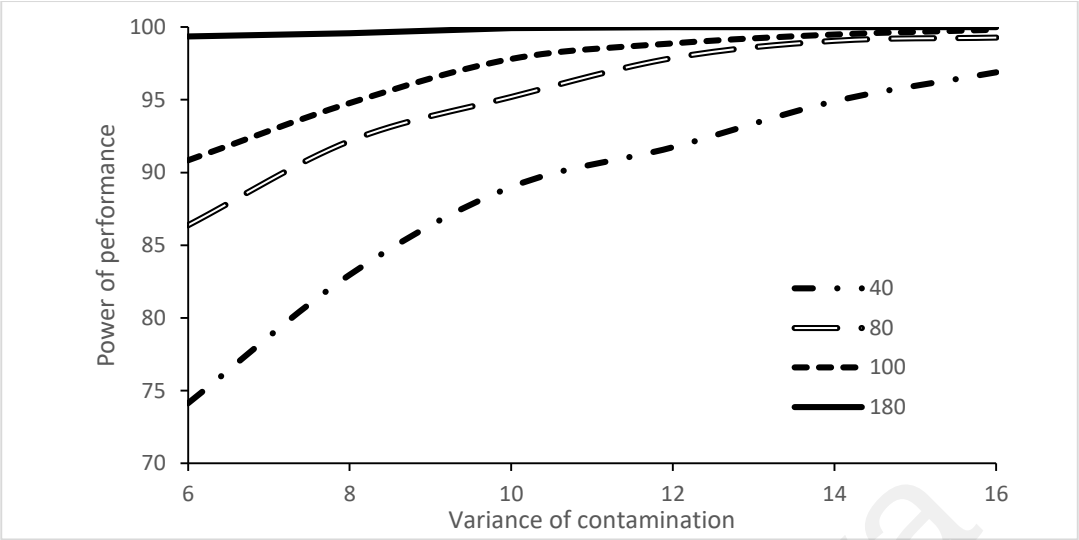

Figure 6.8 Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $\tau^2 = 0.2$.

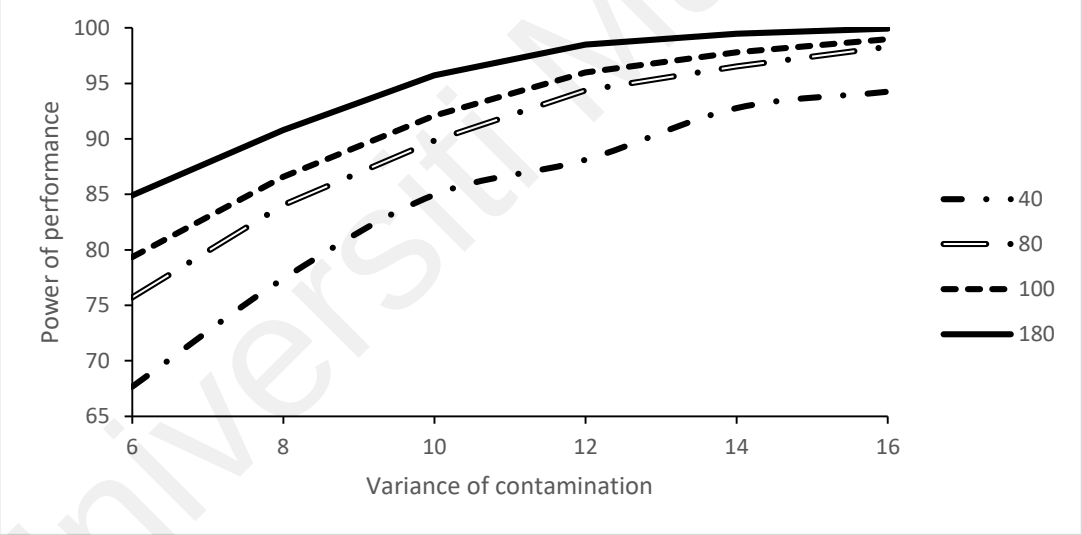

Figure 6.9 Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $\tau^2 = 0.4$.

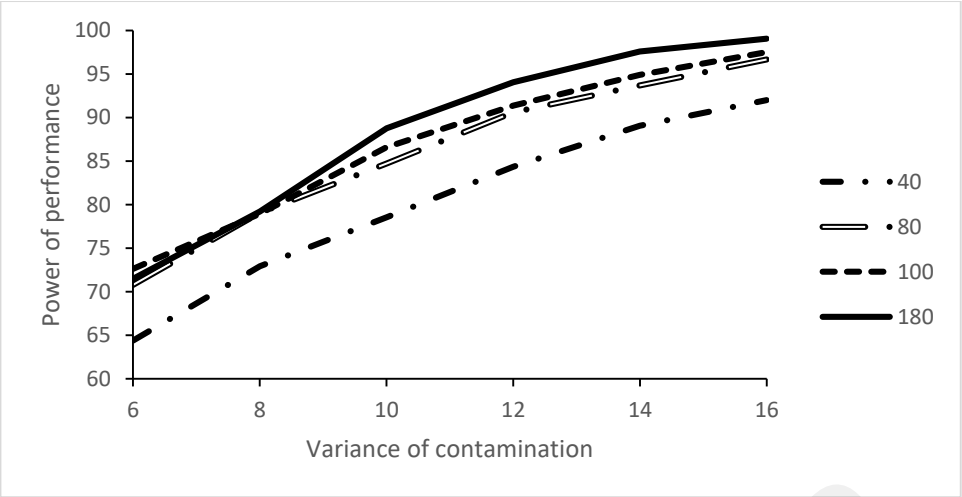

Figure 6.10 Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $\tau^2 = 0.6$

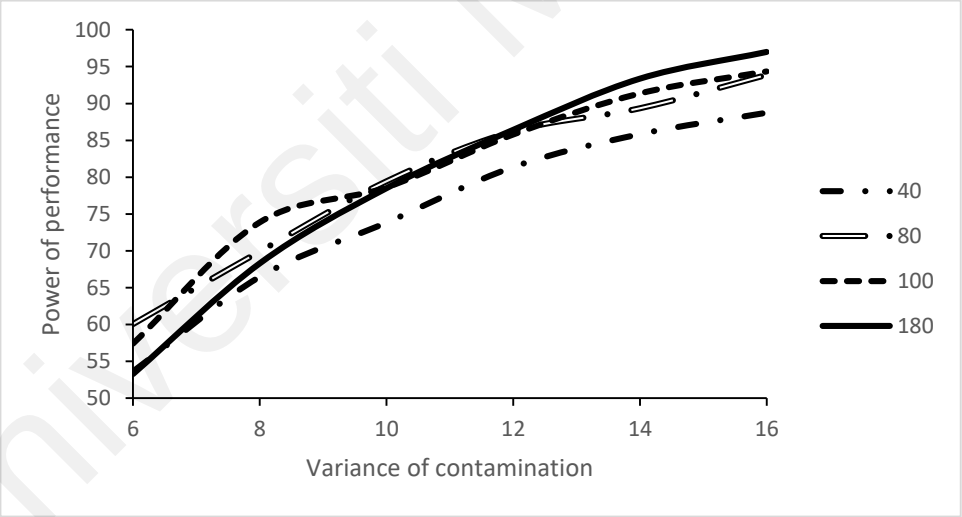

Figure 6.11 Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $\tau^2 = 0.8$

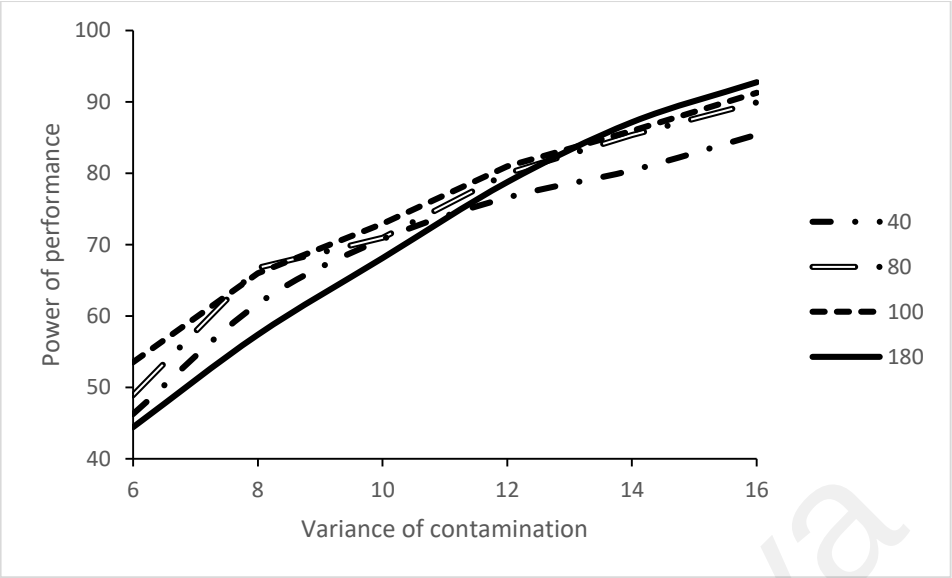

Figure 6.12 Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $\tau^2 = 1.0$.

6.5 Examples

In this section, two data sets to investigate the performance of the cut-off points are considered. The first data set is a simulated data set generated from true values of parameters in balanced replicated linear functional relationship model and the other data sets are taken from systolic blood pressure data. In order to make the relationship as given in equation (4.1), it is assumed that measurement error can occur in both the variables of these two examples.

6.5.1 Simulated Data

For illustration, the data set of size $n = 60$ is considered with six groups, $p = 6$ and each group has 10 observations, $m = 10$. The data is generated from balanced replicated linear functional relationship model given by

$$Y_i = \alpha + \beta X_i, x_{ij} = X_i + \delta_{ij} \text{ and } y_{ij} = Y_i + \varepsilon_{ij}$$

for $i = 1,2,3,4,5,6$ and $j = 1,2,\cdots,10$ and $n = p \times m = 6 \times 10 = 60$. Without loss of generality, the parameters, $\alpha = 0, \beta = 1, \sigma^2 = 1$ and $\tau^2 = 0.4$ is set. Then, the scatterplot of the simulated data is shown in Figure 6.13.

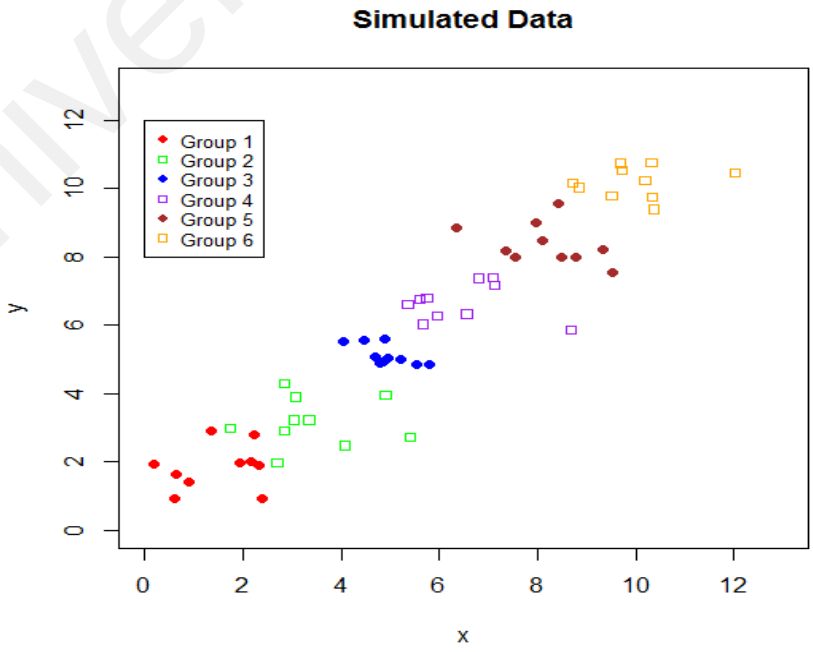

Figure 6.13 The scatter plot for the simulated data

Next, a randomly contamination for the observation, namely at the 11th observation where the contamination is taken from $N(0,12)$ is done. This changed the original point (4.91,3.96) to (4.91, 10.72). The scatterplot of the simulated data sets which includes the contaminated observation are presented in Figure 6.14.

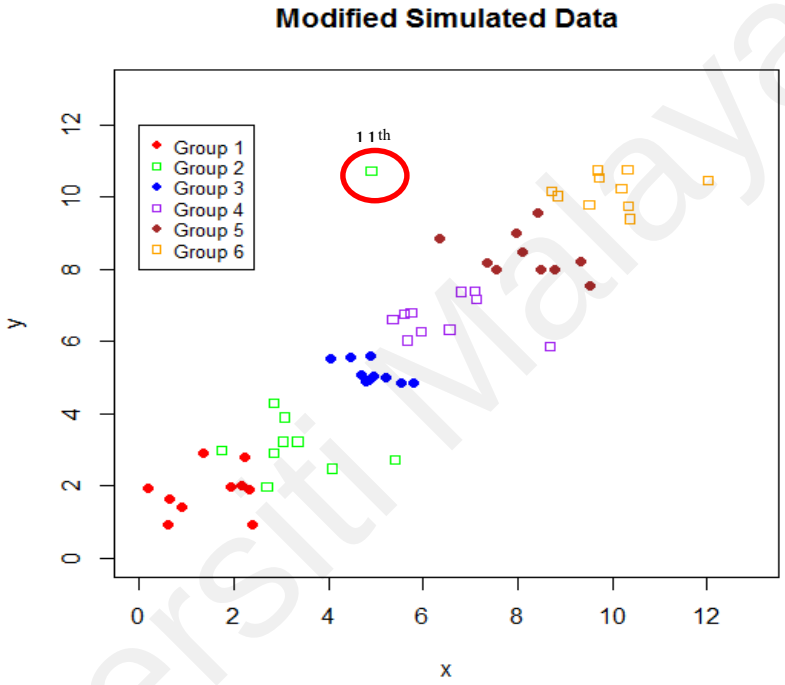

Figure 6.14 The scatter plot for the modified simulated data

From Figure 6.14, it can be seen the 11th observation is slightly far from other observations but still cannot be ascertained as an outlier. The next step is to calculate the *COVRATIO* statistic to the data to determine if there is any outlier. The *COVRATIO* statistic for each value is calculated and the results are given in Table 6.6. Based on the formulation as given in Table 6.5, the cut-off point for $n = 60$ is calculated at 5% significant level and the value 1.118 is obtained. From Table 6.6 and Figure 6.15, it clearly shows that the *COVRATIO* value for 11th observation is 17.9076 which exceeds

the cut-off points of 1.1176. Hence, it can be concluded that the cut-off points correctly identify the 11th observation as an outlier.

Table 6.6 The values for $\left|COVRATIO_{(-i)} - 1\right|$ for the simulated data, $n = 60$.

Index	$\left\|\boldsymbol{COVRATIO_{(-i)} - 1}\right\|$	**Index**	$\left\|\boldsymbol{COVRATIO_{(-i)} - 1}\right\|$	**Index**	$\left\|\boldsymbol{COVRATIO_{(-i)} - 1}\right\|$
1	0.7917	21	0.0227	41	0.3154
2	0.8072	22	0.0653	42	0.4368
3	0.7991	23	0.0689	43	0.3719
4	0.8043	24	0.0658	44	0.3968
5	0.7994	25	0.0667	45	0.4445
6	0.8013	26	0.0552	46	0.4250
7	0.8066	27	0.0199	47	0.3866
8	0.8022	28	0.0658	48	0.4261
9	0.7847	29	0.0721	49	0.4324
10	0.7849	30	0.0442	50	0.4446
11	**17.9076**	31	0.0203	51	0.8019
12	0.2375	32	0.0125	52	0.7912
13	0.3865	33	0.0298	53	0.8069
14	0.3822	34	0.3682	54	0.7985
15	0.4023	35	0.0102	55	0.8063
16	0.2039	36	0.0068	56	0.8065
17	0.3544	37	0.0055	57	0.7954
18	0.2468	38	0.0017	58	0.8057
19	0.3928	39	0.0482	59	0.7570
20	0.3214	40	0.0227	60	0.7977

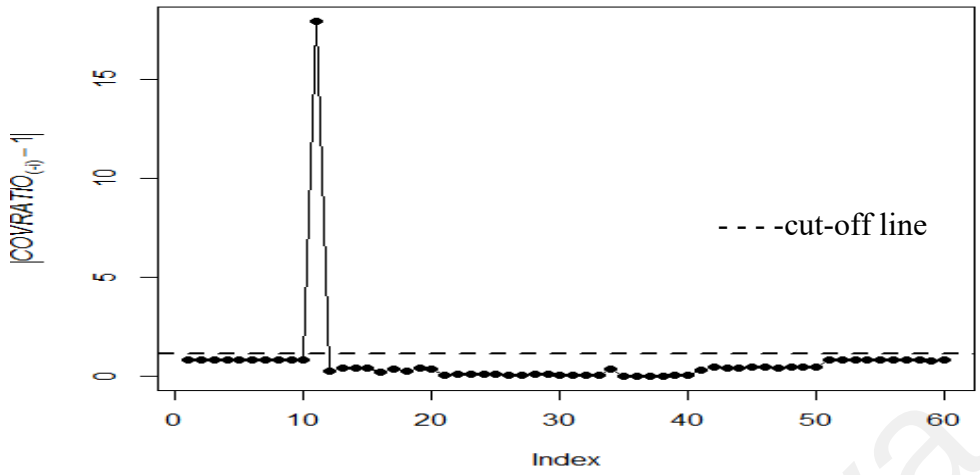

Figure 6.15 Graph of $\left|COVRATIO_{(-i)} - 1\right|$ for simulation data, $n = 60$

Table 6.7 shows the value of parameter estimates and the standard deviation of the modified data in comparison of the original data. As expected, the standard deviation is much higher when the data has a single outlier. This illustrates that when a single outlier presents in the data set, the parameter estimates is affected. Thus, detecting an outlier in the data is important, otherwise the value of the parameter estimates becomes unreliable.

Table 6.7 Parameter estimation and standard deviation for simulated data

Parameter	Original Data (No Outlier)		Modified Data (Single Outlier)	
	Estimates	Standard Deviation	Estimates	Standard Deviation
$\hat{\alpha}$	0.0281	0.3387	0.0507	0.4664
$\hat{\beta}$	1.0169	0.0526	1.0310	0.0727
$\hat{\sigma}^2$	1.0276	0.1876	0.9934	0.1814
$\hat{\tau}^2$	0.3465	0.0633	1.4585	0.2663

6.5.2 Systolic Blood Pressure Data

As another illustration, the subsample of the original dataset containing 30 observations from Bland and Altman (1999) is considered. The data set measures the systolic blood pressure which simultaneous measurements were made by two experienced observers denoted as J and R. It is assumed that measurement error can occur in both the variables Y_i and X_i. In this case, there are 10 groups (or subjects) and each groups have three sets of readings that were made in quick succession. The data can be modelled by

$$Y_i = \alpha + \beta X_i$$

for $i = 1,2,\cdots,10$ and $j = 1,2,3$ and $n = p \times m = 10 \times 3 = 30$. Since there is no outlier in the original data, a random outlier is inserted, i.e. at 14^{th} observation, by following Kim (2000) and Imon and Hadi (2008).

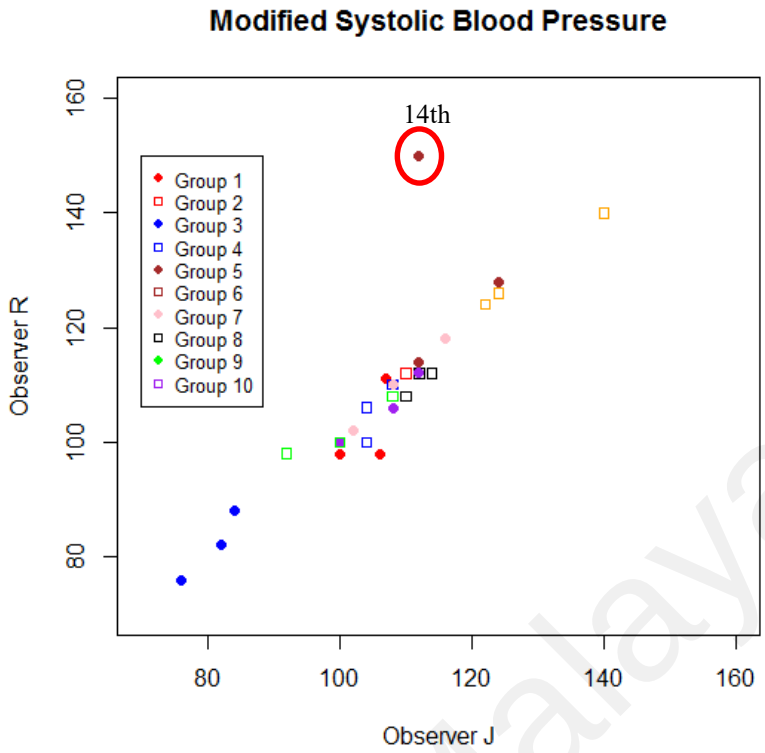

Figure 6.16 The scatter plot for the modified real data

Again, the *COVRATIO* statistic for each value is calculated and the results are given in Table 6.8. Based on the formulation in Table 6.5, the cut-off point for $n = 30$ is calculated, in particular the general formula $y = 9.6293n^{-0.526} = 1.609$ is obtained at 5% significant level.

Table 6.8 The values for $\left|COVRATIO_{(-i)} - 1\right|$ for the real data, $n = 30$

Index	$\left\|\boldsymbol{COVRATIO}_{(-i)} - \mathbf{1}\right\|$	Index	$\left\|\boldsymbol{COVRATIO}_{(-i)} - \mathbf{1}\right\|$	Index	$\left\|\boldsymbol{COVRATIO}_{(-i)} - \mathbf{1}\right\|$
1	0.1705	11	0.1109	21	0.4654
2	0.0687	12	0.0174	22	0.0112
3	0.0478	13	0.3156	23	0.0427
4	0.0003	**14**	**6.3228**	24	0.0604
5	0.0179	15	0.4537	25	0.4483
6	0.0093	16	0.8611	26	0.0132
7	0.9085	17	0.4227	27	0.2504
8	0.9466	18	0.8655	28	0.0333
9	0.9409	19	0.6611	29	0.3529
10	0.0379	20	0.0093	30	0.1852

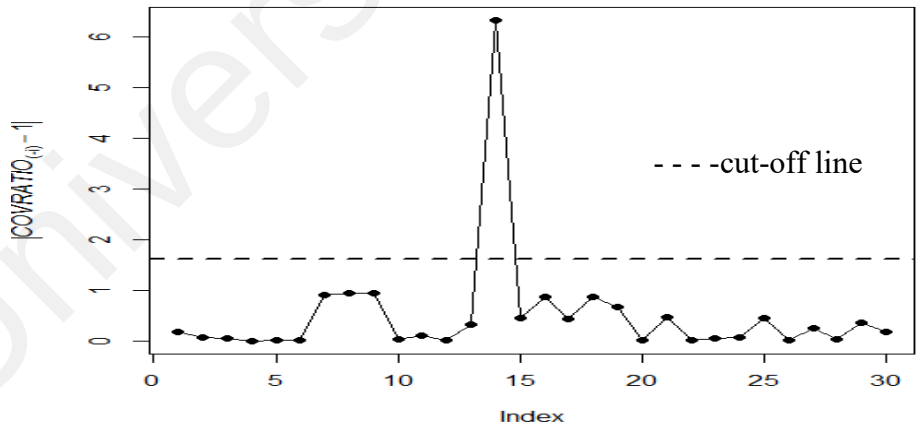

Figure 6.17 Graph of $\left|COVRATIO_{(-i)} - 1\right|$ for real data, $n = 30$

Table 6.8 and Figure 6.17, it is observed that the value of $\left|COVRATIO_{(-i)} - 1\right|$ for the 14th observation is 6.3328, which exceeds the cut-off point value of 1.609 at 5%

significant level. To conclude, based on the calculated cut-off point and the *COVRATIO* statistic used, the method has successfully identified observation 14 as the outlier.

Table 6.9 shows the value of parameter estimates and the standard deviation of the modified data in comparison of the original data. Similar to previous example, the standard deviation of the parameter estimates is much higher when the data has a single outlier. Again, this shows that the value of the parameter estimates is affected when a single outlier presence in the data set. This example illustrates the importance of detecting an outlier in the data. It is crucial since it has an impact on the parameter estimates, thus making them unreliable.

Table 6.9 Parameter estimation and standard deviation for Systolic Blood Pressure data

Parameter	Original Data (No Outlier)		Modified Data (Single Outlier)	
	Estimates	Standard Deviation	Estimates	Standard Deviation
$\hat{\alpha}$	-1.7265	12.3323	-19.4640	16.3882
$\hat{\beta}$	1.0210	0.1145	1.1982	0.1522
$\hat{\sigma}^2$	22.9484	5.9253	25.01363	6.4585
$\hat{\tau}^2$	27.6816	7.1474	51.1897	13.2171

6.6 Summary and Conclusion

In a balanced replicated linear functional relationship model, an outlier detection method has been proposed since the method outlier identification for replicated linear functional relationship model has not been done. A test statistic based on the *COVRATIO* statistic is developed for balanced replicated linear functional relationship model by modifying the covariance function used in the formulation. From simulation studies, the cut-off functions are determined for various significant levels namely 1%, 5% and 10% respectively. Once the cut-off equations are determined, it is important to run the power of performance where it is examined through the simulation studies by looking at the behavior at different level of significant levels, sample size and τ^2 values.

The simulation results suggest that the determined cut-off functions perform well based on the power performance tests. One can use *COVRATIO* statistic to identify outlier in balanced replicated linear functional relationship model with the cut-off functions. To illustrate, two data sets namely simulated and real data set are used. It can be seen from a simulated data and a real dataset where the proposed method can detect the outlier that has been placed randomly in the dataset. It can be concluded that the modified *COVRATIO* statistic can be used in balanced replicated linear functional relationship model. The novelty of the study is that the proposed method is easy to use and can be applied to balanced replicated linear functional relationship model. In summary, it provides a solution for researchers working on a balanced replicated linear functional relationship model not only as a data cleaning and outlier detection tool but also as an important step in all data analysis.

CHAPTER 7: REPLICATING DATA IN LINEAR FUNCTIONAL RELATIONSHIP MODEL USING CLUSTERING ANALYSIS

7.1 Introduction

In this chapter, an effective grouping approach based on clustering is proposed to address the fifth objective of the study. The motivation of this to find a solution to address the problem of unidentifiability in linear functional relationship model. One possible solution is to identify groups from unreplicated data. By grouping the data, the groups formed can be used to estimate the parameters using replicated linear functional relationship model. The replicated linear functional relationship model in this chapter is based on unbalanced observations in each group. In this way, one can overcome the unidentifiability problem in linear functional relationship model. This means, one can estimate the errors variances independently without making any assumption on the ratio of the error variances. Section 7.2 describes in detail the analysis of the agglomerative hierarchical clustering by using the Euclidean distance as the similarity measure. Section 7.3 described the replicated linear functional relationship model by considering different elements or observations in each groups. This is followed by a simulation study in Section 7.4 to test the performance of this clustering algorithm in the linear functional relationship model. Section 7.5 discusses on the result of the simulation study. Simulated data and real data examples are given in Section 7.6. Finally, conclusion and summary of the whole chapter are given in Section 7.7.

7.2 Clustering Methods

In Chapter 4, all the parameters in replicated linear functional relationship model namely the intercept α, the slope β, the error variances σ^2 and τ^2 and the incidental parameters X_i, can be estimated without making any assumption on the ratio of error variances $\lambda = \frac{\tau^2}{\sigma^2}$. By using replicated linear functional relationship model, the unidentifiability problem can be avoided and thus can estimate all the parameters independently. The basic procedures to group the unreplicated data have been discussed in Section 4.4. Nevertheless, the basic grouping techniques in Chapter 4 is too general and need to be explored further. Without having any knowledge how to transform the unreplicated data to replicated data, this chapter proposes some guidelines as to how to group the data using the methods of clustering. Therefore, by transforming the data into replicated data, it can aid the process to estimate all the parameters in linear functional relationship model.

Cluster analysis groups individuals or objects into clusters such that objects in the same cluster are more similar to one another than objects in other clusters. Using the basic idea of cluster analysis, the aim of grouping unreplicated data is to classify different subsets of observations that are almost replicates or almost similar to one another. According to Sebert et al. (1998), similarity measure is needed in order to group the variables or items into their own group. There are four types of similarity measure but the most common used as similarity measure is the distance measure (Blashfield & Aldenderfer, 1978). Euclidean Distance is one type under distance measure and commonly used because it can easily be applied. It can be defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{7.1}$$

where d_{ij} is the distance between i and j, and x_{ik} is the value of the k^{th} variable for the i^{th} observation and x_{jk} is the value of the k^{th} variable for the j^{th} observation (Wang et al., 2005). As an illustration, consider the data in Table 7.1 below:

Table 7.1 Datasets to illustrates Euclidean as a similarity distance

Observations	x	y
1	0.2642	2.0477
2	1.0035	2.4319
3	4.9371	2.9593
4	3.2046	1.3954
5	3.4962	1.9336
6	3.0475	0.7628
7	1.7571	0.8564

As an example, the calculation of Euclidean distance between observation 1 and 2 is given in the following:

$$d_{12} = \sqrt{(0.2642 - 1.0035)^2 + (2.0477 - 2.4319)^2} = 0.8332$$

For the set of data, Euclidean distances are calculated for all the data points using equation in (7.1). Since there are seven observations in this example, the distances obtained are placed in a square matrix with seven rows and seven columns. The distance values, $d_{ij} = d_{ji}$ is written in a similarity matrix in a lower triangular matrix. as shown in Table 7.2.

Table 7.2 The similarity matrix for seven observations

Observation	1	2	3	4	5	6	7
1	0						
2	0.8332	0					
3	4.7609	3.9688	0				
4	3.0119	2.4329	2.3339	0			
5	3.2340	2.5420	1.7687	**0.6121**	0		
6	3.0656	2.6389	2.8974	0.6518	1.2538	0	
7	1.9099	1.7465	3.8124	1.5446	2.0457	1.2938	0

In this study, three different clustering methods from hierarchical cluster analysis (HCA) to cluster the data are considered. The methods are called single-linkage, complete-linkage and average-linkage. Every method has its own procedures and gives different results on grouping. Single linkage method uses the smallest dissimilarity between a point in the first cluster and a point in a second cluster. It is also known as nearest neighbour clustering and one of the most famous of the hierarchical techniques. Complete-linkage method acts in which the similarity of the inter-object is based on the maximum distance between objects in two clusters. The complete linkage also known as furthest neighbour. While for average-linkage method, it represents similarity as the

average distance from all objects in one cluster to all objects in another and also known as the weighted pair-group method. The illustration of the single-linkage, complete-linkage and average linkage can be seen in Figure 7.1.

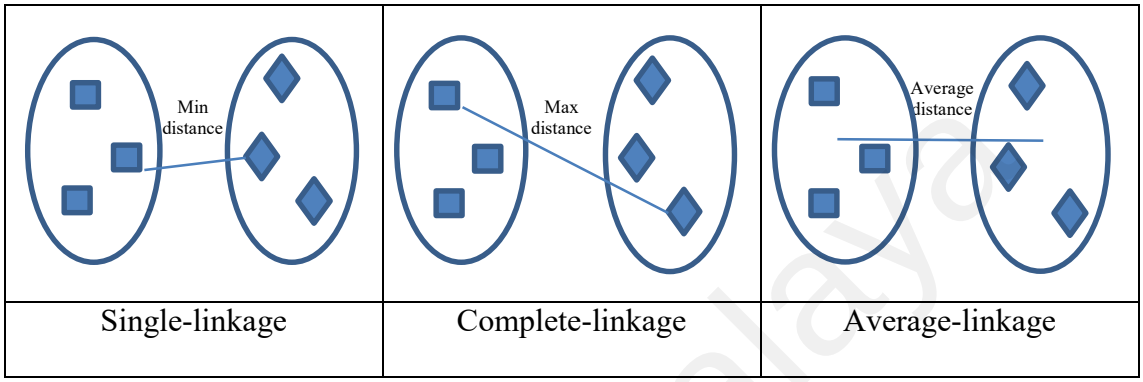

Figure 7.1 Illustration of three selected linkage methods

As an illustration for single-linkage algorithm, in Table 7.2 , there are seven clusters or groups, with one element in each cluster. Then, the similarity measure i.e. the distance matrix using Euclidean distance as in (7.1) is calculated between all possible pairs of cluster. By using single-linkage algorithm, the pair of clusters with the smallest distance is merged. From Table 7.2, the smallest distance is seen in the observation 5 and the observation 4 which is 0.6121. In this step, the observation 5 and the observation 4 is merged, then the row 5 and the column 4 in the similarity matrix are deleted. This process will be repeated until all clusters are merged together where all the observations are combined together in one cluster or the termination condition, for example, the number of groups is set.

The general steps for agglomerative hierarchical clustering algorithm is proposed and explained below:

Step 1: Start with N cluster each containing a single observations, $D = \{d_{ik}\}$.

Step 2: Obtain the distance matrix for the most similar or nearest pairs of clusters. Let the distance between "most similar" clusters U and W be d_{UW}.

Step 3: Merge clusters U and W together. Label the newly formed cluster as UW.

Step 4: Update the entries in the distance matrix by deleting the rows and columns corresponding to clusters U and W. Add a row and column giving the distances between cluster U and W and the remaining clusters.

Step 5: Repeat step 2-4 until only a single cluster remains in the end.

"The most similar" in step 2 is referring to a distance using different agglomerative clustering algorithm namely the single-linkage, complete-linkage and average-linkage. The results of hierarchical clustering methods can be displayed in a form of a dendogram or cluster tree diagram. The height of the dendogram is the distance between clusters. As an illustration, the dendogram based on single-linkage method as described before is shown in Figure 7.2. The dendogram or the cluster tree can be cut depends on how many groups that one intends to have. In this case, it is proposed that the rule or the termination condition will be based on the size of group. Here, the size of the grouped are same but each group will have different elements or observations. This is called as an equal and unbalanced replicated. As mentioned earlier, different linkage methods will result to different clusters or different elements in each group.

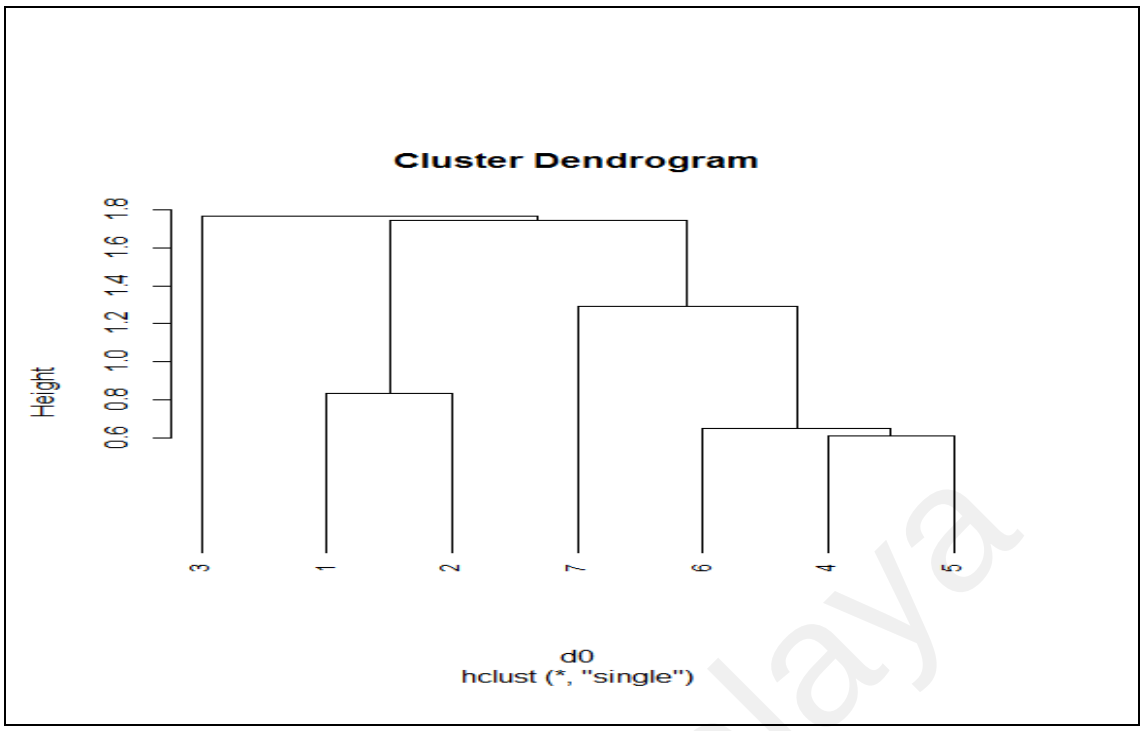

Figure 7.2 Illustration of dendogram based on single-linkage method

The R programming is used to formulate the single-linkage, complete-linkage and average-linkage algorithm as in Figure 7.3. The command line d0=dist(dt_sc, method=’euclidean’), h1=hclust(d0,method='average') and cut_avg <- cutree(h1,k=5), are used where they refer to the similarity measure used is the Euclidean distance, the linkage method used is the average linkage method and the number of cluster groups is 5. The cutree() function is used for the desired number of group or cluster or it can be used as the termination condition.

```
dt_sc <- as.data.frame(scale(dta))      #scale the data
d0=dist(dt_sc, method=’euclidean’)  #calculate the distance using Euclidean Distance
h1=hclust(d0, method='average')      #hierarchical clustering using average-linkage
                                      method
cut_avg <- cutree (h1,k=5)           #the number of cluster/group is 5 (termination
                                      condition)
```

Figure 7.3 The command in R programming for agglomerative hierarchical clustering

7.3 The Replicated Linear Functional Relationship Model

From previous section, each of observation is grouped according to the similarity measures. The size of group is used as the termination condition. Although the number of group is fixed for each clustering methods, the number of observations in each group will not be balanced. In other words, the sizes of group are equal but the observations or elements in each group are not the same. The estimated parameters of linear functional relationship model using the maximum likelihood estimation is obtained for each clustering methods namely the complete-linkage, the average-linkage and the single-linkage algorithm. Recall from Chapter 2, there may be replicated observations of X_i and Y_i occurring in p group. A linear relationship between X_i and Y_i is given by

$$Y_i = \alpha + \beta X_i \tag{7.2}$$

$$\text{and } x_{ij} = X_i + \delta_{ij} \text{ and } y_{ik} = Y_i + \varepsilon_{ik} \tag{7.3}$$

for $i = 1,2,\cdots,p$, $j = 1,2,\cdots,m_i$ and $k = 1,2,\cdots,n_i$ and the errors terms are δ_{ij} and ε_{ik} follow normal distribution with mean zero and variance σ^2 and τ^2 respectively namely $\delta_{ij} \sim N(0,\sigma^2)$ and $\varepsilon_{ik} \sim N(0,\tau^2)$. This implies that

i) both errors have mean 0, that is $E(\delta_{ij}) = 0$ and $E(\varepsilon_{ik}) = 0$, where $i = 1,2,\ldots,p$ and $j = 1,2,\cdots,m_i$ and $k = 1,2,\cdots,n_i$.

ii) both errors have constant but different variance, that is $Var(\delta_{ij}) = \sigma^2$ and $Var(\varepsilon_{ik}) = \tau^2$, where $i = 1,2,\ldots,p$ and $j = 1,2,\cdots,m_i$ and $k = 1,2,\cdots,n_i$.

In this case, the elements in each groups may not be equal or in other words, the replicated linear functional relationship model has p groups but different elements or observations, m_i, and n_i in each groups. This is called equal and unbalanced replicates. The log likelihood function for this model can be expressed as

$$\begin{aligned} &\log L\left(\alpha, \beta, \sigma^2, \tau^2, X_1, \ldots, X_p; x_{ij}, y_{ik}\right) \\ &= -\frac{1}{2}\left(\sum_{i=1}^{p} m_i + \sum_{i=1}^{p} n_i\right)\log 2\pi \\ &\quad -\frac{1}{2}\left(\sum_{i=1}^{p} m_i \log \sigma^2 + \sum_{i=1}^{p} n_i \log \tau^2\right) \\ &-\frac{1}{2}\left\{\sum_{i=1}^{p}\sum_{j=1}^{m_i}\frac{\left(x_{ij} - X_i\right)^2}{\sigma^2} + \sum_{i=1}^{p}\sum_{k=1}^{n_i}\frac{(y_{ik} - \alpha - \beta X_i)^2}{\tau^2}\right\} \end{aligned} \tag{7.4}$$

There are $(p + 4)$ parameters to be estimated and may be obtained by differentiating the log likelihood function as given in (7.4) with respect to $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2$ and $\hat{X}_i$ respectively and equating to zero (Barnett, 1970). Thus, the parameters are obtained in this order given by

$$\hat{X}_i = \frac{1}{\hat{\Delta}}\left\{\frac{m_i \bar{x}_{i.}}{\hat{\sigma}^2} + \frac{n_i \hat{\beta}}{\hat{\tau}^2}(\bar{y}_{i.} - \hat{\alpha})\right\}, \tag{7.5}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{m_i}\left(x_{ij} - \hat{X}_i\right)^2}{\sum_{i=1}^{p} m_i}, \tag{7.6}$$

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{p}\sum_{k=1}^{n_i}\left(y_{ik} - \hat{\alpha} - \hat{\beta}\hat{X}_i\right)^2}{\sum_{i=1}^{p} n_i}, \tag{7.7}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{p} n_i\left(\bar{y}_{i.} - \hat{\beta}\hat{X}_i\right)}{\sum_{i=1}^{p} n_i}, \tag{7.8}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{p} m_i\hat{X}_i(\bar{y}_{i.} - \hat{\alpha})}{\sum_{i=1}^{p} m_i\hat{X}_i^2} \tag{7.9}$$

where $\bar{x}_{i.} = \frac{\sum_{j=1}^{m_i} x_{ij}}{m_i}$ and $\bar{y}_{i.} = \frac{\sum_{k=1}^{n_i} y_{ik}}{n_i}$ are the sample means for each group and $\hat{\Delta}_i = \frac{m_i}{\hat{\sigma}^2} + \frac{n_i\hat{\beta}^2}{\hat{\tau}^2}$.

From (7.6) to (7.9), there are no closed form available as these equations are dependent on $\hat{X}_i$ as given in equation (7.5). Therefore, these estimates can be obtained iteratively by using parameters from unreplicated linear functional relationship as starting values. The iteration will continue until one or all parameters converge.

7.4 Simulation Studies

A simulation studies is performed to determine which clustering algorithm gives the best estimates for the slope parameter. The idea is to group the data using different clustering method and then find the slope estimates using the maximum likelihood estimation method for replicated linear functional relationship model. The data are

generated from unreplicated linear functional relationship model. Recall from Chapter 2, the model is given by:

$$Y_i = 1 + X_i, \quad x_i = X_i + \delta_i, \quad y_i = Y_i + \varepsilon_i \tag{7.10}$$

where $X_i = 10\frac{i}{n}$ and the error terms δ_i and ε_i are assumed to be mutually independent and normally distributed random variables, $\delta_i \sim N(0,0.1)$ and $\varepsilon_i \sim N(0,0.1)$. Different sample size $n = 50,80,100,132$ and 180 are chosen to create different situations. Without loss of generality, the value $\alpha = 0$, $\beta = 1, \tau^2 = 4$ are set and use different values of $\lambda = \frac{\tau^2}{\sigma^2}$ whereby the value of $\lambda = 0.8, 1$ and 1.3 respectively. This will create three different situations namely $\sigma^2 < \tau^2$, $\sigma^2 = \tau^2$ and $\sigma^2 > \tau^2$.

From unreplicated data, three clustering methods namely average-linkage, complete-linkage and simple-linkage are used to divide the observations into different group or cluster based on their similarity measure. After the observations are grouped together, the replicated linear functional relationship model is used to estimate the value of the slope parameter by using the maximum likelihood estimation method. As the baseline for comparisons, the slope parameter from unreplicated linear functional relationship model when λ is assumed to be known and is equal to 1 is used, that is $\lambda = 1$. For each estimate, the observed mean square error over 5000 simulations is calculated. The performance of clustering methods in grouping the observations are examined by observing the mean square error (MSE) of the slope parameter and is given by:

$$\text{MSE} = \frac{1}{s}\sum\left(\widehat{w}_j - w\right)^2$$

where w be a generic term for the slope parameter and s is the number of simulation.

7.5 Results and Discussions

The simulation results based on three different situations are shown in Table 7.3, Table 7.4 and Table 7.5 respectively. Table 7.3 to Table 7.5 shows the mean square error of the slopes estimates using three different clustering method with the baseline unreplicated linear functional relationship model. In unreplicated linear functional relationship model, the assumption for the ratio of error variances must be known and usually is equal to 1 in order to estimate the parameters, in this case the slope parameter. However, this is not the case for replicated linear functional relationship model.

Table 7.3 Mean square error for slope parameter when $\sigma^2 = \tau^2$

Estimates	Sample Sizes				
	50	80	100	132	180
$\beta_{Average}$	0.02371	0.01464	0.00914	0.01235	0.00642
$\beta_{Complete}$	0.02477	0.01481	0.00920	0.01242	0.00644
β_{Single}	0.04318	0.02428	0.01459	0.02036	0.00920
$\beta_{Unreplicate}$	0.02634	0.01520	0.00937	0.01266	0.00657

From Table 7.3, when the error variances are the same, $\sigma^2 = \tau^2$, or when the ratio of the error variances $\lambda = 1$, the mean square error for the slope parameter when using

average-linkage and complete-linkage methods are much smaller compared to unreplicate linear functional relationship model. Clearly, it is observed that the mean square error using single-linkage method is much higher than the other estimates. It is also observed that the mean square errors are getting smaller when the sample sizes increase from 50 to 180 for each slope estimates.

Table 7.4 Mean square error for slope parameter when $\sigma^2 < \tau^2$

Estimates	Sample Sizes				
	50	80	100	132	180
$\beta_{Average}$	0.02641	0.01754	0.01524	0.01223	0.00937
$\beta_{Complete}$	0.02667	0.01754	0.01534	0.01230	0.00945
β_{Single}	0.05270	0.03254	0.02752	0.02164	0.01441
$\beta_{Unreplicate}$	0.02894	0.01822	0.01572	0.01261	0.00967

Table 7.5 Mean square error for slope parameter when $\sigma^2 > \tau^2$

Estimates	Sample Sizes				
	50	80	100	132	180
$\beta_{Average}$	0.02611	0.01736	0.01506	0.01194	0.00948
$\beta_{Complete}$	0.02668	0.01749	0.01513	0.01196	0.00951
β_{Single}	0.03921	0.02283	0.01907	0.01483	0.01099
$\beta_{Unreplicate}$	0.02876	0.01785	0.01538	0.01216	0.00966

From Table 7.4 and Table 7.5, different values of λ i.e. when $\lambda > 1$ and also when $\lambda < 1$ are used to see the consistency of the slope parameter estimate. From Table 7.4, when $\lambda = 0.8$ in which is less than 1 or when the value $\sigma^2 < \tau^2$, the mean square error for each slopes estimates are decreasing when the sample sizes are increasing. In general, the mean square errors for slope estimates using average-linkage and complete-linkage are quite similar and much smaller than the mean square error for slope parameter using unreplicated model. However, the mean square error for single-linkage is much bigger compared to other three estimates. It is also observed the same pattern when $\lambda > 1$ which in this case equal to 1.3 by looking at Table 7.5.

The simulation results suggest that the slope parameter can be estimated using clustering method. By considering different clustering methods, different elements are grouped together depending on their similarity measures. Although the elements in each group are not equal and unbalanced, the slope parameter using replicated linear functional relationship model can be estimated. By comparing to the unreplicated linear functional relationship model, one can infer that using the average-linkage and the complete-linkage yields a reasonable estimates of the slope parameter. The mean square errors for three different situations namely when $\lambda = 1$, $\lambda < 1$ and $\lambda > 1$ give consistent estimates of the value of the slope parameter. The simulation studies suggest a solution to form groupings so that issues of unidentifiability can be addressed. The average-linkage and complete-linkage both show promising potential to do this grouping.

7.6 Examples

In this section, two data sets to investigate the performance of grouping approach using clustering analysis are considered. The first data set is a simulated data set generated from true values of parameter and the second data set is a data set known as Fat Mass Measurement Data. It is assumed that measurement error can occur in both the variables of these two examples.

7.6.1 Simulated Data

Simulated data is used to illustrate the use of clustering method in grouping the observations from unreplicated data to replicated data. For illustration, data set of size $n = 50$ is considered where it is generated from unreplicated linear functional relationship model by setting the parameters, $\alpha = 0, \beta = 1, \tau^2 = 4$ and $\lambda = \frac{\tau^2}{\sigma^2} = 0.8$. Then, the model is given by:

$$Y_i = \alpha + \beta X_i, \quad x_i = X_i + \delta_i, \quad y_i = Y_i + \varepsilon_i \tag{7.11}$$

where $X_i = 10\frac{i}{n}$ and the error terms are δ_i and ε_i follow normal distribution with mean zero and variance σ^2 and τ^2 respectively namely $\delta_i \sim N(0, \sigma^2)$ and $\varepsilon_i \sim N(0, \tau^2)$.

Each observation is defined according to their similarity (dissimilarity) based on clustering methods namely the complete-linkage, average-linkage and single-linkage in

order to cluster the observations into 5 groups. Thus, every clustering method have 5 groups, Group 1 until Group 5 and each group have different observations or elements. As mentioned earlier, this is called balanced and unequal replicates as shown in Table 7.6.

Table 7.6 Number of elements using different clustering method for simulated data

Clustering Method	Group				
	1	2	3	4	5
Complete	14	16	15	3	2
Average	16	13	16	3	2
Single	45	1	1	2	1

Table 7.6 shows the data set and which group the observations belong to. The termination condition is set to five groups, $p = 5$. The number of elements or observations in each groups are different using different clustering methods. The observations in each group can be seen clearly in cluster dendogram and cluster plot for complete-linkage, average-linkage and single linkage method as shown in Figure 7.4, Figure 7.5 and Figure 7.6 respectively. From Figure 7.4 and Figure 7.5, the elements or observations in each group for complete-linkage and the average-linkage are almost the same except the 4^{th}, the 21^{st} and the 32^{nd} observations, unlike single-linkage method. Details on the observations for each can be found in APPENDIX J. Hence, the same estimates for the slope parameter using complete-linkage and average-linkage is expected when using replicated linear functional relationship model.

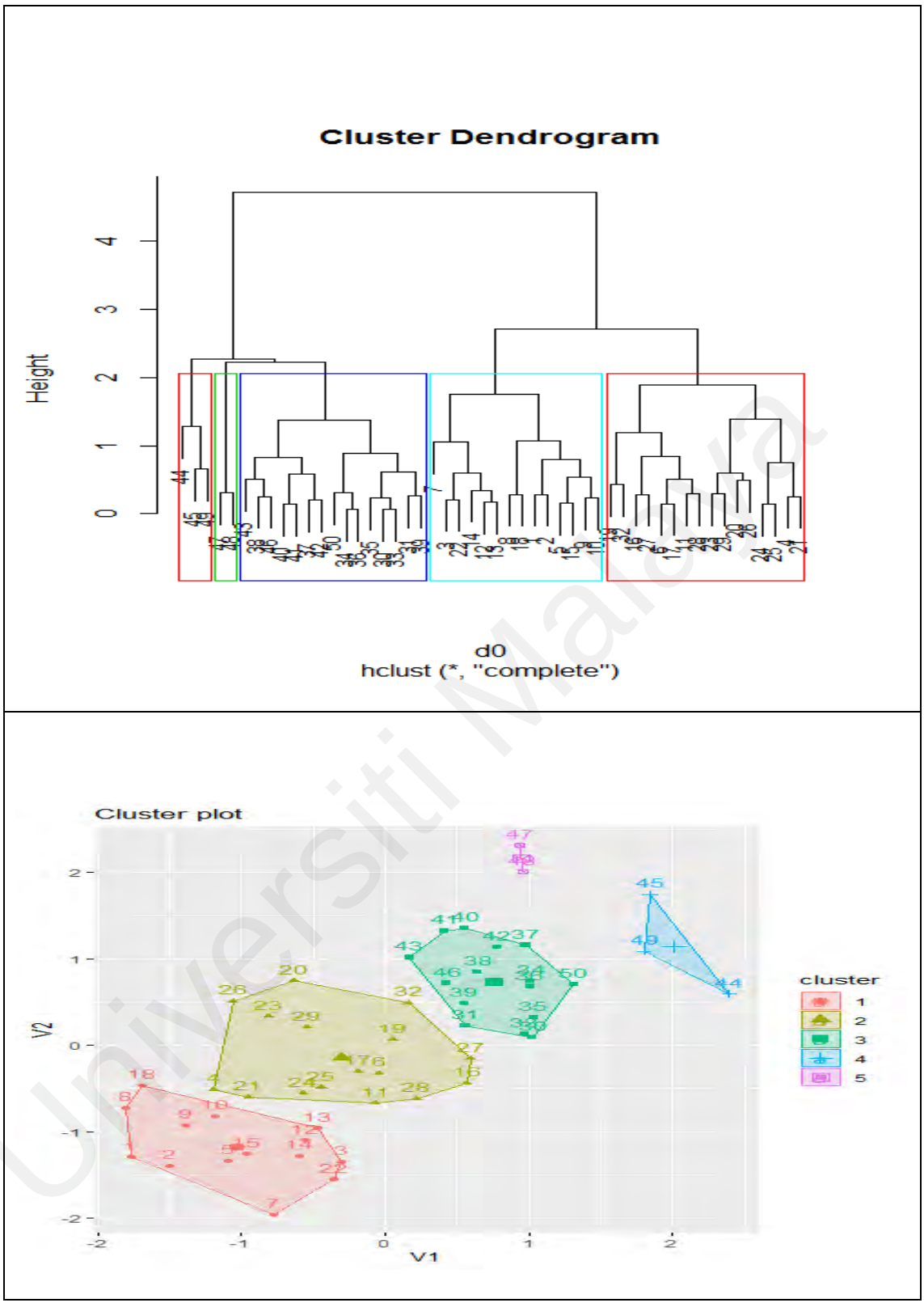

Figure 7.4 Dendogram and cluster plot using complete-linkage method for simulated data

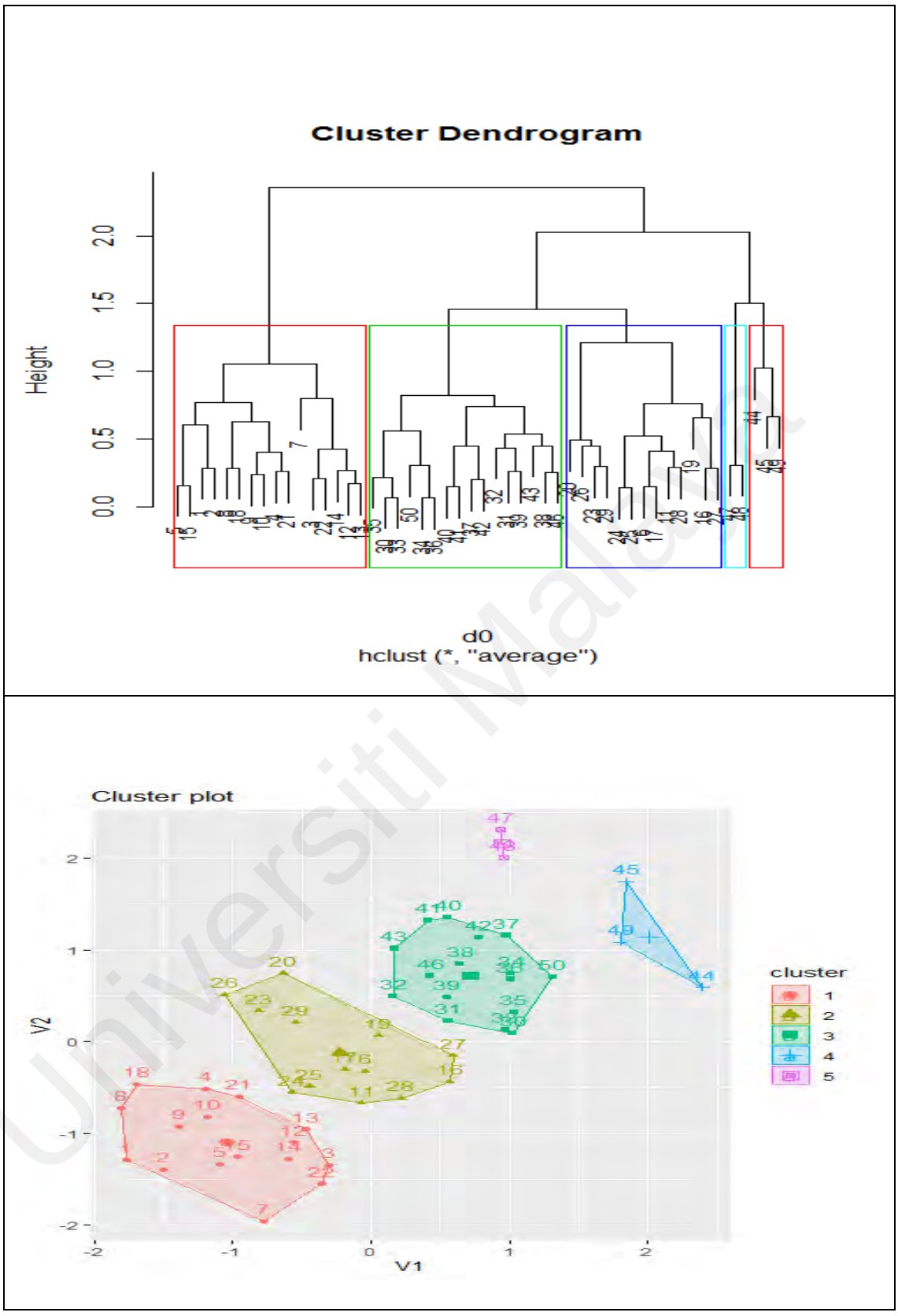

Figure 7.5 Dendogram and cluster plot using average-linkage method for simulated data

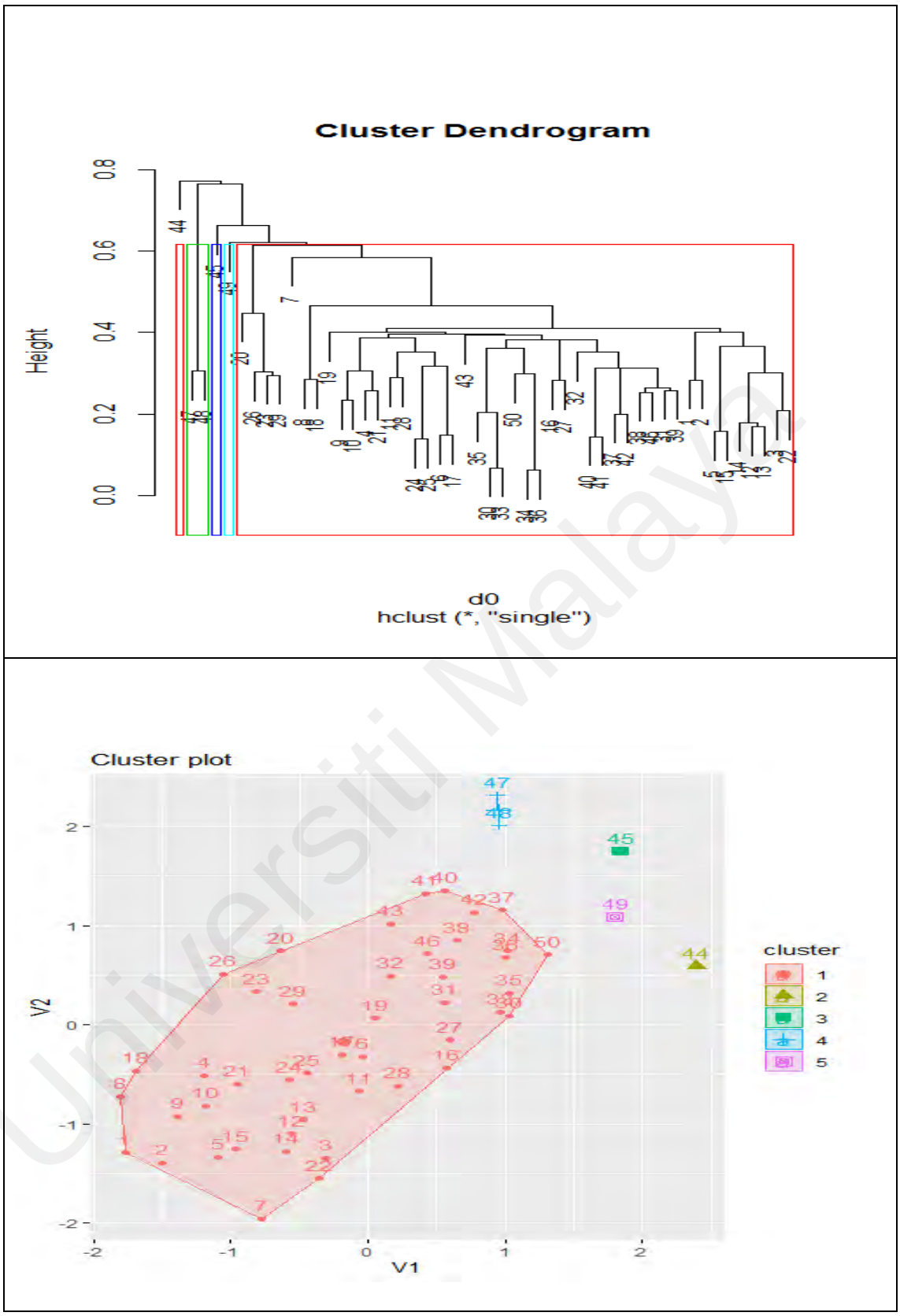

Figure 7.6 Dendogram and cluster plot using single-linkage method for simulated data

From Figure 7.4 to Figure 7.6, the observation is grouped together according to their similarity measures. By setting the termination condition, in this case the size of group, $p = 5$, the observations are grouped into 5 clusters. Although the number of groups is fixed for each clustering methods, the number of observations or elements in each group are not the same in each group. After each observation are grouped to their cluster by the methods described, the estimated parameters of linear functional relationship model using maximum likelihood estimation method are obtained. The result for the parameter estimates are given in Table 7.7.

Table 7.7 Parameter Estimates for simulated data

Clustering Method	Parameter Estimates				
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	λ
Complete	0.035	1.037	2.876	2.545	0.88
Average	0.035	1.035	2.558	2.504	0.98
Single	0.035	1.034	8.825	9.289	1.05
Unreplicated	0.035	1.035	3.993	-	1.00

From Table 7.7, by using three different clustering method, it is obvious that one can estimate all the parameters using replicated linear functional relationship model. In order to measure how well they perform, it is necessary to compare the value of the estimates with unreplicated linear functional relationship model as the baseline. It can be seen that the estimated value of the intercept parameter using three clustering methods are the same with the baseline, unreplicated model. The estimated slope parameter using the complete-linkage method and single-linkage method are 1.037 and 1.034 respectively. The estimated slope parameter using the average-linkage has the same value as the

unreplicated model which is 1.035. while the slope parameter using the clustering method are not much different than slope parameter estimate using unreplicated model.

Recall that in unreplicated linear functional relationship model, the assumption about the ratio of error variance is set which usually equal to one. However, in this simulated data set, the ratio of error variance is set at 0.8, $\lambda = 0.8$. By looking at Table 7.7, it can be seen that the ratio of error variance, λ, the complete linkage gives values 0.88 which is near to the true value. The ratio of error variance from the average linkage and single-linkage are approximately equal to one. Furthermore, both error variances, σ^2 and τ^2 can be estimated as compared to the unreplicated linear functional relationship model without making any assumption about the ratio of error variances. Thus, using clustering method to group the data, one can estimate all the parameters including both error variances using replicated linear functional relationship model. By comparing to the unreplicated linear functional relationship model, it can be inferred that using the complete-linkage provide reasonable estimates of the parameters in linear functional relationship model. Thus, the simulated data suggest that the complete-linkage can be used for groupings to obtain estimate of parameters in linear functional relationship model.

7.6.2 Fat Mass Measurement Data

The dataset is taken from Goran et. al (1996) that consists of 96 observations. This dataset can be considered as unreplicated data because there is only single x and y observation for each level of i as suggested by Hussin et al. (2005) namely the skinfold thickness (ST), x_i and bioelectrical resistance (BR), y_i. In unreplicated linear functional relationship model, one needs an assumption on ratio of error variances, λ, to estimate the

parameters namely the intercept, α, the slope, β and the error variance, σ^2. However, in the absence of knowledge on ratio of error variances, it is necessary to group the data into 8 groups or clusters and use maximum likelihood estimation method for balanced replicated linear functional relationship model to estimate all parameters, in this case, the intercept, α, the slope, β and two error variances, σ^2 and also τ^2. Thus, every clustering method have 8 groups whereas each group have different observations or elements. This is called balanced and unequal replicates as shown in Table 7.8.

Table 7.8 Number of elements using different clustering method for Fat Mass Measurements data

Clustering Method	Group							
	1	2	3	4	5	6	7	8
Complete	57	9	2	20	2	1	3	2
Average	78	2	2	1	3	6	2	2
Single	88	2	1	1	1	1	1	1

Table 7.8 shows the data set and which group the observations belong to. The termination condition is set to five groups, $p = 8$. From Table 7.8, the number of elements or observations in each groups are different using different clustering methods. The observations in each group can be seen clearly in cluster dendogram and cluster plot for complete-linkage, average-linkage and single linkage method as displayed in Figure 7.7, Figure 7.8 and Figure 7.9 respectively. From Figure 7.7, Figure 7.8 and Figure 7.9 the elements or observations in each group for complete-linkage, the average-linkage and the single-linkage are different. Interesting to note that 31^{st} observation is outside the group for all three clustering methods. The 31^{st} observation could be considered a potential outlier. However, this needs to further investigation and could be investigated in future works.

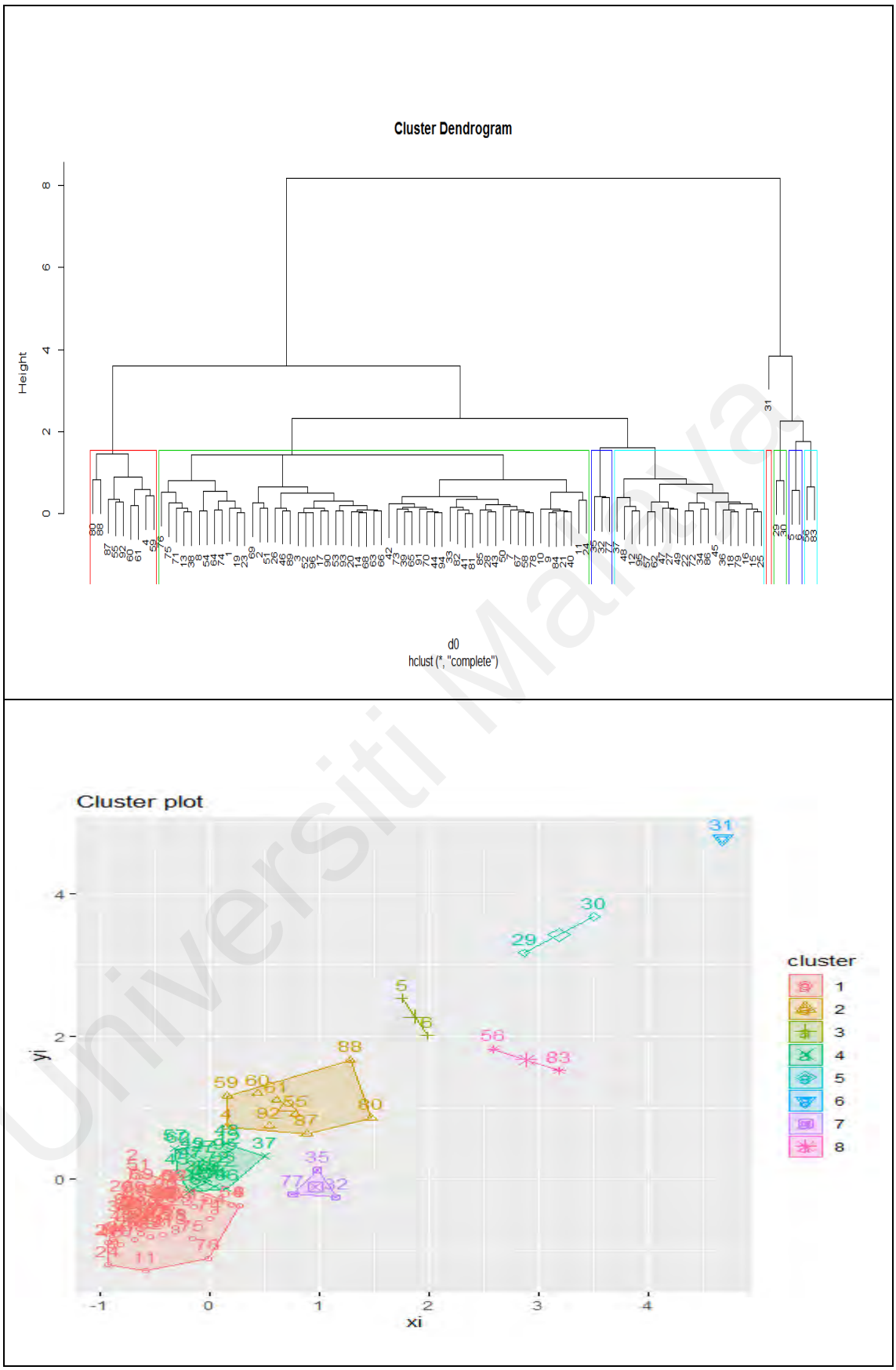

Figure 7.7 Dendogram and cluster plot using complete linkage method for Fat Mass Measurements data

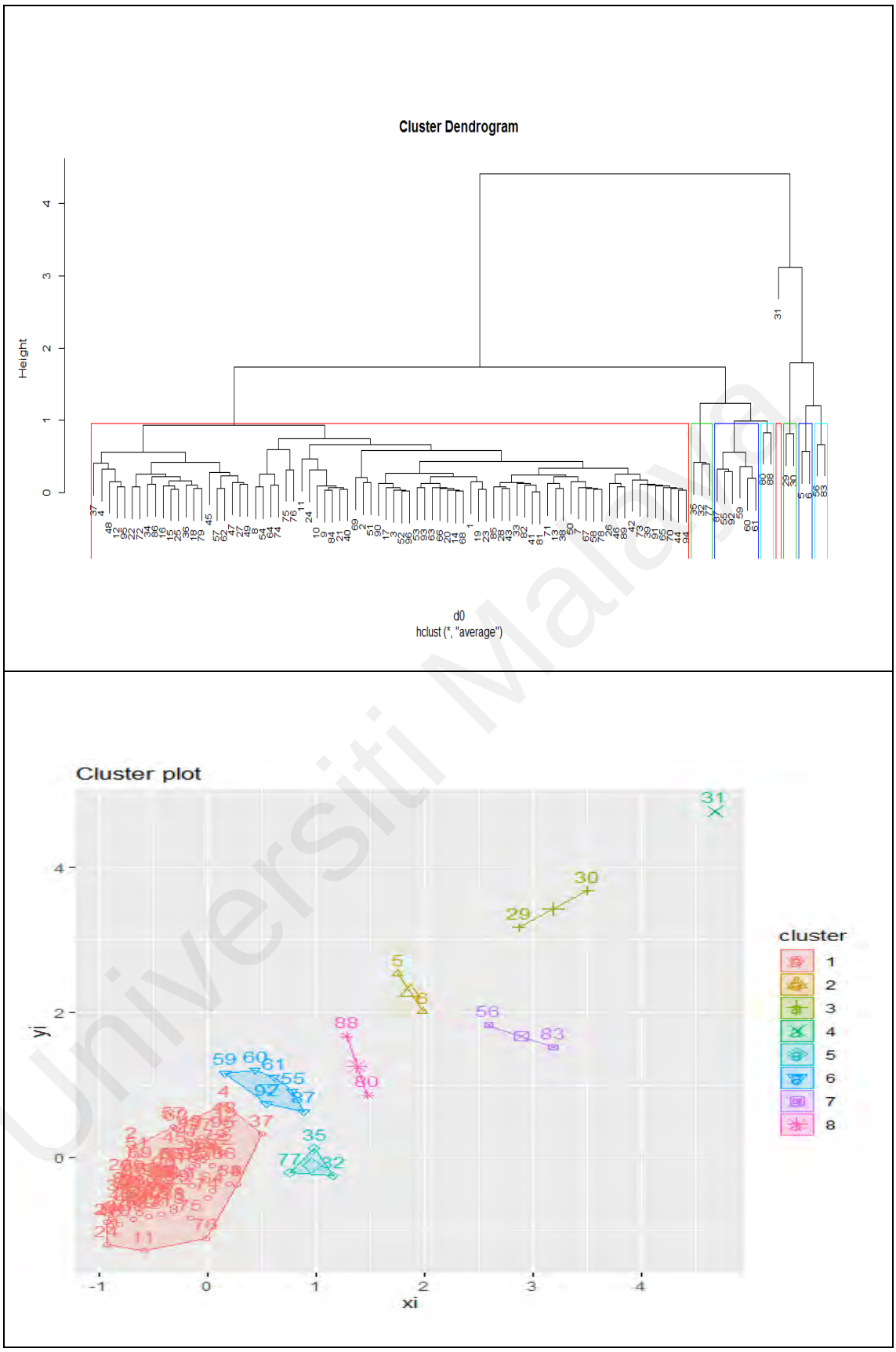

Figure 7.8 Dendogram and cluster plot using average linkage method for Fat Mass Measurements data

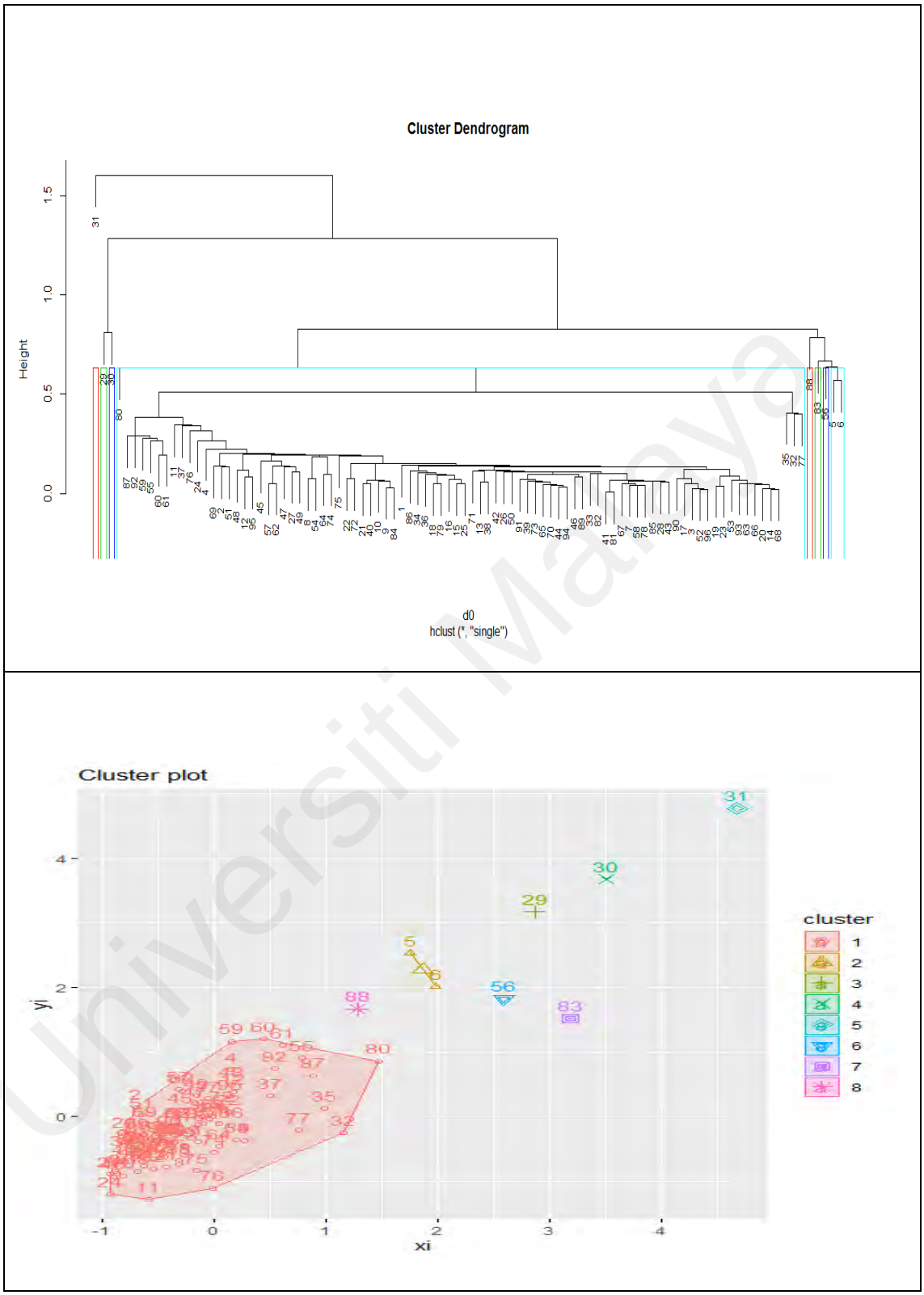

Figure 7.9 Dendogram and cluster plot using single linkage method for Fat Mass Measurements data

From Figure 7.7 to Figure 7.9, it can be seen that each observation is grouped according to the similarity measures. Although the number of groups is fixed, the number of observations or elements are not balanced in each group. After each observation are grouped to their cluster by the method described earlier, the next step is to estimate the parameters of linear functional relationship model using maximum likelihood estimation method for replicated linear functional relationship model. The results for the parameter estimates are given in Table 7.9.

Table 7.9 Parameter Estimates for Fat Mass Measurement Data

Clustering Method	Parameter Estimates				
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	λ
Complete	0.079	1.099	0.966	1.013	1.05
Average	0.079	1.095	1.198	1.941	1.62
Single	0.079	1.096	2.319	3.087	1.33
Unreplicated	0.079	1.099	1.082	-	1.00

Table 7.9 display the estimates of all the parameters using three clustering methods. Also displayed in the table are the estimates of the unreplicated linear functional relationship model which serve as a baseline and for comparison. It can be seen that the value of the intercept parameter, $\hat{\alpha}$, is the same for all three clustering method while the values of the slope, $\hat{\beta}$, is not much different when comparing with the parameters in unreplicated linear functional relationship model. The assumption on the ratio of error variance which usually is equal to one is a must in the unreplicated linear functional relationship model to avoid the unidentifiability problem. From the three clustering methods considered, the ratio of error variance using the complete linkage is

approximately equal to one. For the average-linkage and single-linkage, the ratio of error variances is more than one. Furthermore, the estimate for both error variances, σ^2 and τ^2 are compared with unreplicated linear functional relationship model without making any assumption about the ratio of error variances. In other words, using clustering method to group the data, one can estimate all the parameters including both error variances using replicated linear functional relationship model. Based on the error variance, complete-linkage has the smallest value, indicating the superiority of the method as opposed to average-linkage and single-linkage. In short, it can be concluded that using the complete-linkage yield a reasonable estimates of the parameters as opposed to unreplicated linear functional relationship model.

7.7 Summary and Conclusions

In this chapter, the motivation is to propose a solution in forming groupings to the observations to estimate the parameters when only unreplicated data is available. Using this technique, the beauty is that one does not have to make any assumptions about the ratio of error variances to enable us to estimate the slope parameter. Three grouping techniques using clustering analysis are proposed from unreplicated data. The three different hierarchical clustering methods considered are the single-linkage, the complete-linkage and average-linkage method. Each clustering method has their own strength in calculating the distance for each observation.

Based on the simulation study, this new approach gives us a smaller value of the mean square error as compared with the unreplicated linear functional relationship model with the assumption of the equal variances, or in other words, the ratio of error variances is

equal to one, $\lambda = 1$. By comparing to the unreplicated linear functional relationship model, one could say that using the average-linkage or the complete-linkage yield a reasonable estimates of the slope parameter. Results from data examples also show that one can obtain relatively good estimates of the parameters in particular the slope parameter and also the estimates of ratio of error variance, λ. Based on the datasets, it can be concluded that using the complete-linkage, it can provide a reasonable estimates of the parameters. When comparing the estimates obtained based on this grouping method and by assuming the true value of the ratio of error variance parameters is known, it is found that the value is very close to each other. The novelty of the proposed techniques is that it allows us to overcome the unidentifiability problem in linear functional relationship model and thus, the assumption on the ratio of the error variances are no longer needed unlike in unreplicated linear functional relationship model. By using this technique, it allows us to identify groups of observations and estimate the ratio of error variances separately. Findings from this study will have implications where it not only leads to new knowledge discovery but a scientific progress to providing solutions to the problems faced when modelling linear functional relationship model.

CHAPTER 8: CONCLUSION AND FUTHER WORKS

8.1 Conclusion and summary

This chapter summaries all the findings that have been obtained from this study. This study addresses some problems in linear functional relationship model in particular the parameter estimation, outlier detection and grouping techniques. The linear functional relationship model is chosen due its close resemblance to the linear regression model. The unidentifiability problem in linear functional relationship model makes it difficult to estimate all the parameters and hence warrants further investigations to solve the problem. In order to address this, the linear functional relationship model can be categorized as unreplicated linear functional relationship model where the assumption of ratio of error variance is needed and replicated linear functional relationship model where the observations can be grouped together or made available.

The first problem regarding the parameter estimation can be solved using the first objective. The initial analysis is on the slope parameter in unreplicated linear functional relationship model with an assumption that the error variance is equal $\sigma^2 = \tau^2$ or in other words, the ratio of error variance is equal to one, $\lambda = 1$. The purpose is to propose a robust estimate of the slope parameter and compare with the traditional method namely the maximum likelihood estimation method. The novel approach of modified maximum likelihood estimation estimators by substituting the non-robust elements with robust components is driven by the problem of maximum likelihood estimation estimators which are very sensitive to the outliers. Several cases are considered regarding the distribution namely the symmetric and non-symmetric distribution and also the percentages of outliers

in the dataset. The performance of the proposed method is evaluated and compared with the traditional method, the maximum likelihood estimation method using the estimated bias and the mean square error Empirical evidence from simulation study shows that the proposed method performs well even when 20% outliers are present in the data set. This finding is further illustrated with real data sets.

Still on parameter estimation problem, the parameter estimates of replicated linear functional relationship model is also studied. In a replicated linear functional relationship model, one does not need to make any assumption about the ratio of error variance and can estimate all the parameters independently. To enhance the estimates, the parameters of the balanced replicated model are derived by ensuring the elements in each group are balanced and equal. Estimation are obtained iteratively by taking unreplicated value of parameters as a starting point. Although the closed form of the estimates cannot be obtained, the proposed method is able to obtain the closed form of the variance-covariance matrix using Fisher Information Matrix and partitioned matrix. Based on the empirical results of the simulation study, it is shown that the parameters of balanced replicated linear functional relationship model are unbiased and consistent indicate the adequacy of the proposed model. A practical example is illustrated using real data sets which show the robustness aspects of the parameter estimates as it gives small values of standard deviation.

The third objective focuses on the parameter estimation of the balanced replicated linear functional relationship model in the presence of outlier. Using the traditional method, the parameters in the balanced replicated linear functional relationship model are estimated using the maximum likelihood estimation method which are sensitive to the outliers. To overcome this problem, a robust nonparametric method to estimate the slope parameter is proposed using 20% trimmed mean. Several cases are considered regarding

the distribution namely the symmetric and non-symmetric distribution and also the percentages of outliers in the dataset. The performance of the proposed method is evaluated and compared with the traditional method, the maximum likelihood estimation method using the estimated bias and the mean square error. The estimated bias and the mean square error of the 20% trimmed mean are not affected by outliers regardless of the percentage of outliers present in the dataset. The same results also can be seen in real data examples. Results show that the proposed method provides good estimates in the presence of outliers and the performance measures confirm the superiority of the proposed method.

The fourth objective concentrates on the outlier detection in balanced replicated linear functional relationship model. The *COVRATIO* statistic in detecting the outlier is proposed. In doing so, the cut-off point for 1%, 5% and 10% percentiles are determined. From simulation studies, the general formulation at 1% significant level, 5% significant level and 10% significant level are given by $y = 39.159n^{-0.801}$, $y = 9.6293n^{-0.526}$ and $y = 5.2418n^{-0.407}$ respectively. Any observation that exceeds the cut-off points will be viewed as an outlier. Also, through the simulation study, the power of performance of *COVRATIO* statistic is examined by looking at the behavior at different level of significant levels, sample size and τ^2 values. Based on the results from simulation study, the conclusion can be drawn where the proposed method can detect outlier in a balanced replicated linear functional relationship model.

The final part of the study proposes a grouping approach using hierarchical clustering analysis that convert the unreplicated data to replicated data. Using three different hierarchical clustering methods namely the single-linkage, the complete linkage and the average-linkage, the proposed technique allows us to replicating the observations into group and then use the maximum likelihood estimation to estimate all the parameters without making any assumption on the ratio of error variances. From three clustering

methods, it can be concluded that the average-linkage or the complete linkage can be used in grouping the observations based on the results from simulation study. Again, the method is illustrated using real data for practical applications. Results from simulated data and data example show that one can obtain relatively good estimates of the parameters in particular the slope parameter and also the estimates of ratio of error variance, λ.

8.2 Contributions

This study has contributed to the body of knowledge when working with a linear functional relationship model. In general terms, findings of the study have contributed towards the advancement of statistical analysis. The need for statistical understanding of the linear functional relationship model is really indispensable. In real life, the precision and calibration of the instruments used are hard to obtain. Sometimes, there are some situations where the true variable of interest cannot be obtained directly. These variables involved are subjected to errors. These are some of the vital reasons for studying the errors-in-variables model. In addition, misinterpretation between the linear regression model and the errors-in-variables model particularly in linear functional could be prevented.

The study as a whole addresses pertinent issues in inferential statistics where presence of outliers affects parameter estimation. The study provides solutions to overcome this problem where it contributes to the existing knowledge in ways that brings solution to the problem. The contributions are explained by works done in each chapter. First, when there are outliers in unreplicated linear functional relationship model, the proposed modified

maximum likelihood estimation method provides a robust estimator for the slope parameter. This proposed method remains resistant even when 20% outliers exist in the data. The implication of this findings is that the solution provided contributes to scientific progress in handling outliers in datasets. These outliers could be important because deleting the "good" observations (outliers) may results in underestimating data variability (Maronna et al., 2006).

In replicated linear functional relationship model, an improved estimation using maximum likelihood estimation method called the balanced replicated linear functional relationship model is proposed. The balanced replicated linear functional relationship model can overcome the unidentifiability problem in linear functional relationship model instead of making an assumption on the ratio of error variances. Although the closed form solution is not available, the estimation values of parameters can be done using iteration process and by using suitable initial values. By improving the model, one can derive a step by step asymptotic variance covariance matrix using Fisher Information Matrix and partitioned matrix. Empirical evidence from the simulation study support the proposed method where the parameters obtained are unbiased and consistent.

The study also provided solution when the outliers are present in balanced replicated linear functional relationship model. The proposed nonparametric method can be considered as a new method and uses 20% trimmed mean to estimate the slope parameter in replicated linear functional relationship model. This new method in balanced replicated linear functional relationship model is found to be robust to outliers. In other words, this method is performed well in the existence of the outliers. The beauty of this method is that the nonparametric method does not require the assumption of normality, gives a robust estimate and thus, making it easy to apply.

Another important topic that is addressed in the study is the identification of outlier namely a single outlier in a balanced replicated linear functional relationship model which is a relatively new topic and has not been explored fully. A single outlier will have adverse influence in the dataset and hence will affect the estimation of parameters. The proposed method will provide a technique that can be used in a data quality evaluation process. By deriving the cut-off point using the *COVRATIO* statistic, the proposed method is able to identify a single outlier in a straightforward and a simple way.

The last part of the study is to propose grouping techniques using hierarchical clustering method in obtaining the group of the observations from unreplicated data. This allows us to overcome the unidentifiability problem in linear functional relationship model and thus, the assumption on the ratio of the error variances are no longer needed unlike in unreplicated linear functional relationship model. The hierarchical clustering methods can be used to find the possible group of the observations and estimate the ratio of error variances separately and thus can estimate all the parameters. The novelty of this approach is a scientific progress by providing solutions to problems encountered i.e. the unidentifiability problem in modelling linear functional relationship model.

8.3 Limitation of the Study and Further Works

Despite the contributions that have been discussed in previous section, there are still many areas that can be developed for future works. The scope of the study results in some limitations. One limitation in this study is that the focus is on estimating the slope parameter. This is because the slope parameter could be considered as one of the important parameters to be estimated in order to find a relationship between variables in

many practical applications. In Chapter 3, this research could be extended to estimate other parameters namely the intercept and the error variances and study the performance on these parameters instead. In other words, further research could investigate more on the robust component such as using the trimmed mean and also study the effect of the trimming for the modified maximum likelihood estimation method in unreplicated linear functional relationship model. The same can be said for the nonparametric approach in Chapter 5, where in this study, the scope is on the slope parameter only. In addition, future research on non-zero intercept values and negative slope values can be conducted to investigate the behavior of the estimator.

Moreover, the error term could be investigated further using different distributions such as the heavy-tailed distribution and the extreme distribution, though beta distribution is particularly flexible in modelling different curves including symmetrical, left and right skewed curves. Some researchers used the heavy-tailed distribution on the unreplicated linear functional relationship model instead of replicated model (Tomaya & de Castro, 2018; Duwarahan & Nawarathna, 2022). The study can also look into the extreme distributions for example the Weibull and Gumbel distributions, which are widely used in reliability modeling. Further works can be done on other parameters such as the intercept or the error variances in replicated linear functional relationship model. This work also can be extended to the multiple linear functional relationship model in estimating the parameters.

The study also focuses on estimating the parameters for replicated linear functional relationship model with balanced observations in each group as in Chapter 4. This work can be extended to a general situation where there could be unbalanced observations in each group. Another point is selecting the initial value to start the iteration process which

could be from regression model. Further works also can be done using a nonparametric approach where the observations in each group are unbalanced.

In the identification of outliers in a linear functional relationship model in Chapter 6, this study only examined situations when there is a single outlier at the response variable, *y*. Further works can be extended by putting outliers in the *x* variables or outliers in both *x* and *y* variables. The diagnostic tools in detecting outlier also can be explored by using other methods such as Cook's distance, Difference in fits (DIFFITS), Difference in beta (DFBETA) which have been discussed in the literatures. Furthermore, the outlier detection also can be extended into identifying the multiple outliers in the dataset.

For the grouping approach in the last chapter, one can consider using another clustering method. Some possible methods may include the partitioned clustering method such as PAM and K-Means for grouping the observation. It is also worthy to consider using another distance measure such as Mahalanobis and Manhattan distance in calculating the similarity measure of the data. The termination condition under cluster analysis can also be studied to group the dataset. Furthermore, clustering analysis can be utilized to identify single or multiple outliers in replicated linear functional relationship model. As an illustration, the 31^{st} observation in Fat Mass Measurement data could become as a potential outlier and need to be investigated more.

Additional work could include a wide range of applications in variety of fields. This includes applications based on supervised and unsupervised machine learning. The integration between the errors-in-variables model with the concept of uncertainty quantification in deep learning has been discussed in details by Martin and Elster (2021). As mentioned by Pividori et al., (2022), by using machine learning, one could identify pattern that can be used to group data into their similarity measures. Statistical process control is another application that can be used which includes the EWMA control chart

and the CUSUM control chart, (Golosnoy et al., 2022; Song et al., 2022). The impact of measurement error to detect a process shift is very important key in control chart.

REFERENCES

Abdullah, M. (1995). Detection of Influential Observations in Functional Errors-in-variables Model. *Communications in Statistics - Theory and Methods*, *24*(6), 1585–1595.

Abdullah, M. B. (1989). On Robust Alternatives to the Maximum Likelihood Estimators of a Linear Functional Relationship. *Pertanika*, *12*(1), 89–98.

Abuzaid, A., Mohamed, I., Hussin, A. G., & Rambli, A. (2011). COVRATIO Statistic for Simple Circular Regression Model. *Chiang Mai Journal Science*, *38*(3), 321–330.

Acock, A. C. (2005). Working With Missing Values. *Journal of Marriage and Family*, *67*, 1012–1028.

Adcock, R. J. (1878). A Problem in Least Squares. *Annals of Mathematics*, *5*(2), 53–54.

Al-Nasser, A. D., & Ebrahem, M. A. H. (2005). A New Nonparametric Method for Estimating the Slope of Simple Linear Measurement Error Model in the Presence of Outliers. *Pak. J. Statist*, *21*(3), 265–274.

Anderson, T. W. (1984). Estimating Linear Statistical Relationships. *The Annals of Statistics*, *12*(1), 1–45.

Barker, F., Soh, Y. C., & Evans, R. J. (1988). Properties of the Geometric Mean Functional Relationship. *Biometrics*, *44*(1), 279–281.

Barnett, V. D. (1970). Fitting Straight Lines-The Linear Functional Relationship with Replicated Observations. *Applied Statistics*, *19*(2), 135–144.

Bartlett, M. S. (1949). Fitting a Straight Line When Both Variables are Subject to Error. *Biometrics*, *5*(3), 207–212.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.

Birch, M. W. (1964). A Note on the Maximum Likelihood Estimation of a Linear Structural Relationship. *Journal of the American Statistical Association*, *59*(308), 1175–1178.

Bland, J. M., & Altman, D. G. (1999). Measuring Agreement in Method Comparison Studies. *Statistical Methods in Medical Research*, *8*(2), 135–160.

Blashfield, R. K., & Aldenderfer, M. S. (1978). The Literature on Cluster Analysis. *Multivariate Behavioral Research*, *13*(3), 271–295.

Buonaccorsi, J. P. (2010). *Measurement Error Models, Methods and Applications*. New York: Chapman and Hall.

Carroll, R. J. (1998). Measurement Error in Epidemiologic Studies. In *Encyclopedia of Biostatistics* (Vol. 3, pp. 2491–2519). John Wiley & Sons.

Cateni, S., Colla, V., Vannucci, M., Aramburo, J., & Trevino, A. R. (2008). Outlier Detection Methods for Industrial Applications. *Advances in Robotics, Automation and Control*, 265–282.

Chan, L. K., & Mak, T. K. (1979). Maximum Likelihood Estimation of a Linear Structural Relationship with Replication. *J.R. Statist. Soc. Ser. B*, *41*(2), 263–268.

Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity Analysis In Linear Regression*. New York: John Wiley & Sons.

Cheng, C. L., & Ness, J. W. Van. (1994). On Estimating Linear Relationship When Both Variables are Subject to Errors. *Journal of Royal Statistical Society.Series B (Methodological)*, *56*(1), 167–183.

Cheng, C. L., & Ness, J. W. Van. (1999). *Statistical Regression with Measurement Error*. Arnold; New York: Oxford University Press.

Cook, R. D. (1979). Influential Observations in Linear Regression. *Journal of the American Statistical Association*, *74*(365), 169–174.

Copas, J. B. (1972). The Likelihood Surface in the Linear Functional Relationship Problem. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 274–278.

Dasgupta, S., & Long, P. M. (2005). Performance Guarantees for Hierarchical Clustering. *Journal of Computer and System Sciences*, *70*(4), 555–569.

Deming, W. E. (1931). XI. The Application of Least Squares. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *11*(68), 146–158.

Dent, B. M. (1935). On Observations of Points Connected by a Linear Relation. *Proceedings of the Physical Society*, *47*(1), 92–108.

Di, N. F. M., & Satari, S. Z. (2017). The Effect of Different Distance Measures in Detecting Outliers using Clustering-based Algorithm for Circular Regression Model. *AIP Conference Proceedings*, *1842*.

Doganaksoy, N., & Van Meer, H. (2015). An Application of the Linear Errors-in-Variables Model in Semiconductor Device Performance Assessment. *Quality Engineering*, *27*(4), 500–511.

Dolby, G. R. (1976). The Ultrastructural Relation: A Synthesis of the Functional and Structural Relations. *Biometrika*, *63*(1), 39–50.

Dolby, G. R., Cormack, R. M., & Sinclair, D. F. (1987). On Fitting Bivariate Functional Relationships to Unpaired and Unequally Replicated Data. *Biometrika*, *74*(2), 393–399.

Dolby, G. R., & Lipton, S. (1972). Maximum Likelihood Estimation of the General Nonlinear Functional Relationship with Replicated Observations and Correlated Errors. *Biometrika*, *59*(1), 121–129.

Dorff, M., & Gurland, J. (1961a). Estimation of the Parameters of a Linear Functional Relation. *Journal of the Royal Statistical Society. Series B (Methodological)*, *23*(1), 160–170.

Dorff, M., & Gurland, J. (1961b). Small Sample Behavior of Slope Estimators in a Linear Functional Relation. *Biometrics*, *17*(2), 283–298.

Drion, E. F. (1951). Estimation of the Parameters of a Straight Line and of the Variances of the Variables, If They are Both Subject to Error. *Indagationes Mathematicae (Proceedings)*, *54*, 256–260.

Dunn, G. (2004). *Statistical Evaluation of Measurement Errors* (2nd Ed). London: Arnold.

Durbin, J. (1954). Errors in Variables. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, *22*(1/3), 23–32.

Duwarahan, J., & Nawarathna, L. S. (2022). An Improved Measurement Error Model for Analyzing Unreplicated Method Comparison Data under Asymmetric Heavy-Tailed

Distributions. *Journal of Probability and Statistics*, *2022*, 1–13.

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th Ed). John Wiley & Sons.

Fah, C. Y., Hussin, A. G., & Rijal, O. M. (2007). An Investigation of Causation:The Unreplicated Linear Functional Relationship Model. *Journal of Applied Sciences*, *7*(1), 20–26.

Fekri, M., & Ruiz-Gazen, A. (2004). Robust Weighted Orthogonal Regression in the Errors-in-variables Model. *Journal of Multivariate Analysis*, *88*(1), 89–108.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons.

Geary, R. C. (1942). Inherent Relations between Random Variables. *Proceedings of the Physical Society. Sect A: Mathematical and Physical*, *47*, 63–76.

Geary, R. C. (1943). Relations between Statistics : The General and the Sampling Problem When the Samples Are Large. *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences*, *49*, 177–196.

Geary, R. C. (1949). Determination of Linear Relations between Systematic Parts of Variables with Errors of Observation the Variances of Which Are Unknown. *Econometrica*, *17*(1), 30–58.

Gençay, R., & Gradojevic, N. (2011). Errors-in-variables estimation with wavelets. *Journal of Statistical Computation and Simulation*, *81*(11), 1545–1564.

Ghapor, A. A., Zubairi, Y. Z., Mamun, A. S. M. Al, & Imon, A. H. M. R. (2014). On Detecting Outlier in Simple Linear Functional Relationship Model using COVRATIO Statistic. *Pak. J. Statist*, *30*(1), 129–142.

Ghapor, A. A., Zubairi, Y. Z., Mamun, A. S. M. Al, & Imon, A. H. M. R. (2015). A Robust Nonparametric Slope Estimation in Linear Functional Relationship Model. *Pak. J. Statist*, *31*(3), 339–350.

Gillard, J. W. (2007). *Errors in Variables Regression: What is the Appropriate Model? (doctoral dissertation)*. Retrieved from http://ethos.bl.uk

Gillard, J. W. (2014). Method of Moments Estimation in Linear Regression with Errors in Both Variables. *Communications in Statistics - Theory and Methods*, *43*(15), 3208–3222.

Gillard, J. W., & Iles, T. C. (2006). Variance Covariance Matrices for Linear Regression with Errors in both Variables. In *Cardiff University School of Mathematics Technical Report*. Cardiff, UK.

Golosnoy, V., Hildebrandt, B., Köhler, S., Schmid, W., & Seifert, M. I. (2022). Control Charts for Measurement Error Models. *AStA Advances in Statistical Analysis*, 1–20.

Golub, G. H., & Van Loan, C. F. (1980). An Analysis of the Total Least Squares Problem. *SIAM Journal on Numerical Analysis*, *17*(6), 883–893.

Goran, M. I., Driscoll, P., Johnson, R., Nagy, T. T. R., & Hunter, G. (1996). Cross-calibration of Body-Composition Techniques Against Dual-energy X-ray Absorptiometry in Young Children. *American Journal of Clinical Nutrition*, *63*(3), 299–305.

Graybill, F. A. (1961). *An Introduction to Linear Statistical Models*. New York: McGraw-Hill.

Hajek, J. (1969). *A Course in Nonparametric Statistics*. Holden-Day.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2011). *Robust Statistics The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., & Ostrowski, E. (1994). *A Handbook of Small Data Sets* (First Ed). Chapman and Hall.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons.

Hill, M., & Dixon, W. J. (1982). Robustness in Real Life: A Study of Clinical Laboratory Data. *Biometrics*, *38*(2), 377–396.

Housner, G. W., & Brennan, J. F. (1948). The Estimation of Linear Trends. *The Annals of Mathematical Statistics*, *19*(3), 380–388.

Hussin, A. G. (2004). Numerical Comparisons for Various Estimators of Slope Parameters for Unreplicated Linear Functional Model. *Matematika*, *20*(1), 19–30.

Hussin, A. G. (2005). Approximating Fisher's Information for the Replicated Linear Circular Functional Relationship Model. *Bulletin of the Malaysian Mathematical Sciences Society*, *28*(2), 131–139.

Hussin, A. G., & Abuzaid, A. (2012). Detection of Outliers in Functional Relationship Model for Circular Variables via Complex Form. *Pak. J. Statist*, *28*(2), 205–216.

Hussin, A. G., Abuzaid, A. H., Ibrahim, A. I. ., & Rambli, A. (2013). Detection of Outliers in the Complex Linear Regression Model. *Sains Malaysiana*, *42*(6), 869–874.

Hussin, A. G., Abuzaid, A., Zulkifili, F., & Mohamed, I. (2010). Asymptotic Covariance and Detection of Influential Observations in a Linear Functional Relationship Model for Circular Data with Application to the Measurements of Wind Directions.

ScienceAsia, *36*(3), 249–253.

Hussin, A. G., Fieller, N., & Stillman, E. (2005). Pseudo-replicates in the Linear Circular Functional Relationship Model. *Journal of Applied Sciences*, *5*(1), 138–143.

Ibrahim, S., Rambli, A., Hussin, A. G., & Mohamed, I. (2013). Outlier Detection in a Circular Regression Model Using COVRATIO Statistic. *Communications in Statistics - Simulation and Computation.*, *42*(10), 2272–2280.

Imon, A. H. M. R., & Hadi, A. . (2008). Identification of Multiple Outliers in Logistic Regression. *Communications in Statistics-Theory and Methods*, *37*(11), 1697–1709.

Isogawa, Y. (1985a). A Note on a Linear with Structural Replication. *Journal Japan Statistical Society*, *15*(1), 71–74.

Isogawa, Y. (1985b). Estimating a Multivariate Linear Structural Relationship with Replication. *Journal of the Royal Statistical Society.Series B (Methodological)*, *47*(2), 211–215.

Isogawa, Y. (1992). A Note on Estimating Bivariate Structural Relationships with Unpaired and Unequally Replicated Observations. *Journal Japan Statistical Society*, *22*(2), 193–200.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys*, *31*(3), 1–60.

Johnson, R. A., & Wichern, D. (2015). Multivariate Analysis. *Wiley StatsRef: Statistics Reference Online*, 1–20.

Jolicoeur, P. (1975). Linear Regressions in Fishery Research: Some Comments. *Journal of the Fisheries Research Board of Canada*, *8*(32), 1491–1494.

Jung, K. M. (2007). Least Trimmed Squares Estimator in the Errors-in-Variables Model. *Journal of Applied Statistics*, *34*(3), 331–338.

Jung, Y., Park, H., Du, D. Z., & Drake, B. L. (2003). A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering. *Journal of Global Optimization*, *25*(1), 91–111.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Kendall, M. G. (1952). Regression, Structure and Functional Relationship. Part II. *Biometrika*, *39*(1/2), 96–108.

Kendall, M. G., & Stuart, A. (1979). The Advanced Theory of Statistics. Vol 2: Inference and Relationship. In *London:Griffin* (4th ed., Vol. 2). London: Griffin.

Kim, M. G. (2000). Outliers and Influential Observations in the Structural Errors-in-variables Model. *Journal of Applied Statistics*, *27*(4), 451–460.

Klepper, S., & Leamer, E. E. (1984). Consistent Sets of Estimates for Regressions with Errors in All Variables. *Econometrica*, *52*(1), 163–184.

Koláček, J. (2008). Plug-in Method for Nonparametric Regression. *Computational Statistics*, *23*(1), 63–78.

Koul, H. L., & Song, W. (2008). Regression Model Checking with Berkson Measurement Errors. *Journal of Statistical Planning and Inference*, *138*(6), 1615–1628.

Kummel, C. H. (1879). Reduction of Observation Equations Which Contain More Than One Observed Quantity. *The Analyst*, *6*(4), 97–105.

Lindley, D. V. (1947). Regression Lines and the Linear Functional Relationship. *Supplement to the Journal of the Royal Statistical Society*, *9*(2), 218–244.

Lindley, D. V., & El-Sayyad, G. M. (1968). The Bayesian Estimation of a Linear Functional Relationships. *Journal of the Royal Statistical Society Series B (Methodological)*, *30*(1), 190–202.

Liu, C. A. (2012). A Plug-In Averaging Estimator for Regressions with Heteroskedastic Errors. *Available at SSRN 2138013*.

Madansky, A. (1959). The Fitting of Straight Lines when Both Variables are Subject to Error. *Journal of the American Statistical Association*, *54*(285), 173–205.

Maindonald, J., & Braun, J. (2010). *Data Analysis and Graphics Using R – an Example-Based Approach* (3rd Ed). Cambridge University Press.

Mamun, A. S. M. Al, Hussin, A. G., Zubairi, Y. Z., & Rana, S. (2020). A Modifed Maximum Likelihood Estimator for the Parameters of Linear Structural Relationship Model. *Malaysian Journal of Mathematical Sciences*, *14*(2), 209–220.

Mamun, A. S. M. Al, Zubairi, Y. Z., Hussin, A. G., Imon, A. H. M. R., Rana, S., & Carrasco, J. (2019). Identification of Influential Observation in Linear Structural Relationship Model with Known Slope. *Communications in Statistics - Simulation and Computation*, 1–12.

Markovsky, I., & Van Huffel, S. (2007). Overview of Total Least-squares Methods. *Signal Processing*, *87*(10), 2283–2302.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). Robust statistics: Theory and Methods. In *John Wiley*. John Wiley & Sons.

Martin, J., & Elster, C. (2021). Errors-in-Variables for Deep Learning: Rethinking Aleatoric Uncertainty. *ArXiv Preprint ArXiv 2105.09095*, (2), 1–13.

Milligan, G. W., & Cooper, M. C. (1985). An Examination of Procedures for Determining the Number of Clusters in Data Set. *Psychometrika*, *50*(2), 159–179.

Mirmozaffari, M., Golilarz, N. A., Band, S. S., & Mosavi, A. (2020). *Machine Learning Algorithms Based on an Optimization Model Data Mining Algorithms Based on an Optimization Model*. 1–22.

Mirmozaffari, M., Yazdani, R., Shadkam, E., Tavassoli, L. S., & Massah, R. (2021). VCS and CVS: New Combined Parametric and Non-parametric Operation Research Models. *Sustainable Operations and Computers*, *2*, 36–56.

Moberg, L., & Sundberg, R. (1978). Maximum Likelihood Estimation of a Linear Functional Relationship When One of the Departure Variances is Known. *Scandinavian Journal of Statistics*, *5*(1), 61–64.

Mokhtar, N. A., Badyalina, B., Chang, K. L., Al Mamun, A. S. M., & Zubairi, Y. Z. (2022). Butterworth Wind Direction Statistical Model with Functional Relationship. *Applied Mathematical Sciences*, *16*(12), 565–572.

Mokhtar, N. A., Zubairi, Y. Z., & Hussin, A. G. (2017). Detecting Multiple Outliers in Linear Functional Relationship Model for Circular Variables using Clustering Technique. *AIP Conference Proceedings*, 1–7.

Mokhtar, N. A., Zubairi, Y. Z., Hussin, A. G., & Moslim, N. H. (2019). An Outlier Detection Method for Circular Linear Functional Relationship Model using COVRATIO Statistics. *Malaysian Journal of Science*, *38*(2), 46–54.

Mokhtar, N. A., Zubairi, Y. Z., Hussin, A. G., & Yunus, R. M. (2017). On Parameter Estimation of a Replicated Linear Functional Relationship Model for Circular Variables. *Matematika*, *33*(2), 159–163.

Montfort, K. van. (1989). *Estimating in Structural Models With Non-Normal Distributed Variables: Some Alternative Approaches*. Leiden: DSWO Press.

Montgomery, D. C., Peck, E. A., & Vinning, G. G. (2012). *Introduction to Linear Regression Analysis* (5th Ed). Wiley, New York.

Moran, P. A. P. (1971). Estimating Structural and Functional Relationships. *Journal of Multivariate Analysis*, *1*(2), 232–255.

Murtagh, F., & Contreras, P. (2012). Algorithms for Hierarchical Clustering: An Overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 86–97.

Neyman, J., & Scott, E. L. (1951). On Certain Methods of Estimating the Linear Structural Relation. *The Annals of Mathematical Statistics*, *22*(3), 352–361.

Nurunnabi, A. A. M., Rahmatullah Imon, A. H. M., & Nasser, M. (2011). A Diagnostic Measure for Influential Observations in Linear Regression. *Communications in Statistics - Theory and Methods*, *40*(7), 1169–1183.

O'Driscoll, D., & Ramirez, D. E. (2011). Geometric View of Measurement Errors. *Communications in Statistics - Simulation and Computation*, *40*(9), 1373–1382.

Oosterhoff, J. (1994). Trimmed Mean or Sample Median? *Statistics and Probability Letters*, *20*(5), 401–409.

Pal, M. (1980). Consistent Moment Estimators of Regression Coefficients in the Presence

of Errors in Variables. *Journal of Econometrics*, *14*(3), 349–364.

Partovi Nia, V., & Davison, A. C. (2015). A Simple Model-based Approach to Variable Selection in Classification and Clustering. *Canadian Journal of Statistics*, *43*(2), 157–175.

Pearson, K. (1901). LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Pividori, M., Ritchie, M. D., Milone, D. H., & Greene, C. S. (2022). An Efficient not-only-linear Correlation Coefficient based on Machine Learning. *BioRxiv*, *06*.

Rambli, A., Abuzaid, A., Mohamed, I., & Hussin, A. G. (2016). Procedure for Detecting Outliers in a Circular Regression Model. *PLoS ONE*, *11*(4), 1–10.

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., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons.

Ruppert, D., & Carroll, R. J. (1980). Trimmed Least Squares Estimation in the Linear Model. *Journal of the American Statistical Association*, *75*(372), 828–838.

Satari, S. Z., & Khalif, K. M. N. K. (2020). Review on Outliers Identification Methods for Univariate Circular Biological Data. *Advances in Science, Technology and Engineering Systems*, *5*(2), 95–103.

Satman, M. H. (2013). A New Algorithm for Detecting Outliers in Linear Regression. *International Journal of Statistics and Probability*, *2*(3), 101–109.

Sebert, D. M., Montgomery, D. C., & Rollier, D. A. (1998). A Clustering Algorithm for Identifying Multiple Outliers in Linear Regression. *Computational Statistics & Data Analysis*, *27*(4), 461–484.

Shalabh, Paudel, C. M., & Kumar, N. (2009). Consistent Estimation of Regression Parameters under Replicated Ultrastructural Model with Non-normal Errors. *Journal of Statistical Computation and Simulation*, *79*(3), 251–274.

Sharif, O., Hasan, Z., Fah, C. Y., & Sirdari, M. Z. (2019). Efficiency Analysis by Combination of Frontier Methods: Evidence from Unreplicated Linear Functional Relationship Model. *Business and Economic Horizons*, *15*(1), 107–125.

Shevlyakov, G., & Smirnov, P. (2011). Robust Estimation of the Correlation Coefficient: An Attempt of Survey. *Austrian Journal of Statistics*, *40*(1), 147–156.

Singh, S., Jain, K., & Sharma, S. (2012). Using Stochastic Prior Information in Consistent Estimation of Regression Coefficients in Replicated Measurement Error Model. *Journal of Multivariate Analysis*, *111*, 198–212.

Singh, S., Jain, K., & Sharma, S. (2014). Replicated Measurement Error Model under Exact Linear Restrictions. *Statistical Papers*, *55*(2), 253–274.

Sokal, R. R. (1963). The Principles and Practice of Numerical Taxonomy. *Taxon*, *12*(5), 190–199.

Song, L., He, S., Zhou, P., & Shang, Y. (2022). Empirical likelihood ratio charts for profiles with attribute data and random predictors in the presence of within-profile correlation. *Quality and Reliability Engineering International*, *38*(1), 153–173.

Sprent, P. (1970). The Saddle Point of the Likelihood Surface for a Linear Functional

Relationship. *Journal of the Royal Statistical Society. Series B (Methodological)*, *32*, 432–434.

Sprent, P. (1990). Some History of Functional and Structural Relationships. *Contemporary Mathematics*, *112*, 3–15.

Sprent, P., & Smeeton, N. C. (2016). Applied Nonparametric Statistical Methods. In *CRC Press*.

Stigler, S. M. (1973). The Asymptotic Distribution of the Trimmed Mean. *The Annals of Statistics*, *1*(No. 3), 472–477.

Teissier, G. (1948). La Relation D ' Allometrie sa Signification Statistique et Biologique. *Biometrics*, *4*(1), 14–53.

Theil, H. (1950). A Rank-Invariant Method of Linear and Polynomial Regression Analysis. *Indagationes Mathematicae*, *12*(85), 173.

Tomaya, L. C., & de Castro, M. (2018). A Heteroscedastic Measurement Error Model based on Skew and Heavy-tailed Distributions with Known Error Variances. *Journal of Statistical Computation and Simulation*, *88*(11), 2185–2200.

Tony Cai, T., & Hall, P. (2006). Prediction in Functional Linear Regression. *Annals of Statistics*, *34*(5), 2159–2179.

Van Aelst, S., Wang, X., Zamar, R. H., & Zhu, R. (2006). Linear Grouping using Orthogonal Regression. *Computational Statistics and Data Analysis*, *50*(5), 1287–1312.

Van Huffel, S., Lemmerling, P., & (Eds.). (2002). *Total Least Squares and Errors-in variables Modeling: Analysis, Algorithm and Applications*. Springer Science &

Business Media.

Van Huffel, S., & Vandewalle, J. (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, Philadelphia.

Villegas, C. (1961). Maximum Likelihood Estimation Of A Linear Functional Relationship. *The Annals of Mathematical Statistics*, *32*, 1048–1062.

Wald, A., & Wolfowitz, J. (1940). On a Test Whether Two Samples are from the Same Population. *The Annals of Mathematical Statistics*, *11*(2), 147–162.

Wang, L., Zhang, Y., & Feng, J. (2005). On the Euclidean Distance of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1334–1339.

Warren Liao, T. (2005). Clustering of Time Series Data - A Survey. *Pattern Recognition*, *38*(11), 1857–1874.

Welsh, A. H. (1987). The Trimmed Mean in the Linear Model. *The Annals of Statistics*, *15*(1), 20–36.

Wilcox, R. R. (2005). Trimmed Means. *Encyclopedia of Statistics in Behavioral Science*, *4*, 2066–2067.

Wong, M. Y. (1989). Likelihood Estimation of a Simple Linear Regression Model when Both Variables have Error. *Biometrika*, *76*(1), 141–148.

Xu, R., & Wunsch, D. C. (2010). Clustering Algorithms in Biomedical Research: A Review. *IEEE Reviews in Biomedical Engineering*, *3*, 120–154.

Zamar, R. H. (1989). Robust Estimation in the Errors-in-variables Model. *Biometrika*, *76*(1), 149–160.