PARAMETER ESTIMATION AND OUTLIER DETECTION IN LINEAR FUNCTIONAL RELATIONSHIP MODEL

ADILAH BINTI ABDUL GHAPOR

INSTITUTE OF GRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

PARAMETER ESTIMATION AND OUTLIER DETECTION IN LINEAR FUNCTIONAL RELATIONSHIP MODEL

ADILAH BINTI ABDUL GHAPOR

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

INSTITUTE OF GRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **Adilah binti Abdul Ghapor** (I.C. No: [redacted])

Matric No: **HHC130019**

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

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

Parameter Estimation and Outlier Detection in 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: **3/3/2017**

Subscribed and solemnly declared before,

Witness's Signature Date: **3/3/2017**

Name:

Designation:

ABSTRACT

This research focuses on the parameter estimation, outlier detection and imputation of missing values in a linear functional relationship model (LFRM). This study begins by proposing a robust technique for estimating the slope parameter in LFRM. In particular, the focus is on the non-parametric estimation of the slope parameter and the robustness of this technique is compared with the maximum likelihood estimation and the Al-Nasser and Ebrahem (2005) method. Results of the simulation study suggest that the proposed method performs well in the presence of a small, as well as high, percentage of outliers. Next, this study focuses on outlier detection in LFRM. The *COVRATIO* statistic is proposed to identify a single outlier in LFRM and a simulation study is performed to obtain the cut-off points. The simulation results indicate that the proposed method is suitable to detect a single outlier. As for the multiple outliers, a clustering algorithm is considered and a dendogram to visualise the clustering algorithm is used. Here, a robust stopping rule for the cluster tree base on the median and median absolute deviation (MAD) of the tree heights is proposed. Simulation results show that the proposed method performs well with a small value of masking and swamping, thus implying the suitability of the proposed method. In the final part of the study on the missing value problem in LFRM, the modern imputation techniques, namely the expectation-maximization (EM) algorithm and the expectation-maximization with bootstrapping (EMB) algorithm is proposed. Simulation results show that both methods of imputation are suitable in LFRM, with EMB being superior to EM. The applicability of all the proposed methods is illustrated in real life examples.

ABSTRAK

Kajian ini memberi tumpuan kepada penganggaran parameter, pengesanan data terpencil dan kaedah imputasi untuk nilai lenyap bagi model linear hubungan fungsian (LFRM). Kajian ini dimulakan dengan mencadangkan teknik yang kukuh untuk menganggar kecerunan model linear hubungan fungsian. Khususnya, kajian ini berfokus kepada anggaran kecerunan model menggunakan kaedah tidak berparameter, dan kekukuhan pendekatan ini dibandingkan dengan kaedah kebolehjadian maksimum dan kaedah Al-Nasser dan Ebrahem (2005). Daripada keputusan simulasi, kaedah yang dicadangkan memberi keputusan yang bagus ketika peratusan data terpencil rendah dan tinggi. Seterusnya, kajian ini memberi tumpuan kepada pengesanan data terpencil bagi LFRM. Kaedah mengesan satu data terpencil menggunakan statistik "*COVRATIO*" dicadangkan bagi model LFRM dan simulasi dijalankan untuk memperoleh titik potongan. Keputusan simulasi menunjukkan kaedah yang dicadangkan ini berjaya dalam mengesan satu data terpencil. Apabila wujudnya data terpencil berganda, penggunaan algoritma berkelompok dipertimbangkan serta ilustrasi menggunakan dendogram digunakan. Kaedah yang lebih kukuh dicadangkan untuk nilai potongan bagi pokok kelompok berdasarkan median dan median sisihan mutlak (MAD) bagi ketinggian pokok tersebut. Keputusan simulasi menunjukkan kaedah yang dicadangkan berjaya mengesan data terpencil berganda di dalam sesebuah set data dan menunjukkan prestasi yang bagus dengan nilai "masking" dan "swamping" yang rendah. Bahagian akhir kajian ini mengambil kira nilai lenyap dalam LFRM dan penggantian menggunakan kaedah moden, iaitu kaedah maksima kebarangkalian (EM) dan kaedah maksima kebarangkalian dengan "bootstrap" (EMB) dicadangkan. Keputusan menunjukkan kedua-dua kaedah sesuai digunakan dalam model LFRM, dengan kaedah EMB lebih memuaskan daripada kaedah EM. Penggunaan kesemua kaedah yang dicadangkan ditunjukkan menggunakan contoh data set yang sebenar.

ACKNOWLEDGEMENT

First and foremost, all praises to Allah the Most Merciful and Most Compassionate for giving me the strength and opportunity to complete this doctoral thesis. I would like to express my deepest gratitude to my dedicated supervisor, Associate Professor Dr. Yong Zulina Zubairi and my respectable advisor, Professor Imon Rahmatullah for their advice, motivation, and relentless knowledge sharing throughout my candidature. Their guidance helped me to persevere in this research and complete this thesis. I would also like to acknowledge my helpful research team for the endless support, stimulating discussions, and for the honest and valuable feedback throughout this ups and downs journey. A sincere gratitude goes to University of Malaya and Kementerian Pendidikan Malaysia for the willingness to financially support me to pursue my passion since 2012.

Special thanks to my dear mother and father, Roslinah Mahmood and Abdul Ghapor Hussin for all the known and unknown sacrifices that you both had done to ease this challenging journey. Words cannot express how grateful I am to have the presence of you two in my life. To my mother-in-law and father-in-law, Fatimah Ahmad and Muhamad Yusof Yahya, my siblings; Aimi Nadiah, Amirah, and Amirulafiq as well as my siblings-in-law; Fatasha, Fakhruddin , Eleena, Liyana, Ariff, and Aiman, you have all aided me physically and spiritually and walked hand in hand with me in completing this adventure. To Puan Fatimah Wati and her family, I am grateful for all the help and sacrifices that you have given all these while in taking care of my children while I am away, trying my best to complete this thesis.

For the apples of my eyes; my dear son and daughter, Amjad Sufi and Athifah Safwah, despite the challenges of being a mother throughout this incredible journey, you two have been my huge inspiration and motivation towards accomplishing my studies. Last but not least, I would like to share this memory with my beloved husband, Amirul

Afiq Sufi for his understanding, encouragement, patience and unwavering love that have fuelled me in surviving the experience of being a student in graduate school. Thank you again to all whom I have mentioned and to whom I may miss out, please know that my prayers and utmost thanks will always be with you. May Allah repay all of you justly.

TABLE OF CONTENTS

CHAPTER 3: NONPARAMETRIC ESTIMATION FOR SLOPE OF LINEAR FUNCTIONAL RELATIONSHIP MODEL

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Y	Mathematical variable for a functional relationship model that is linearly related with X
X	Mathematical variable for a functional relationship model that is linearly related with Y
α	Intercept parameter
β	Slope parameter
δ_i	Random error term for the independent variable
ε_i	Random error term for the dependent variable
λ	Ratio of the error concentration parameters in a functional relationship model
σ	Standard error of the model
S	Sum of square
D	Distance
i	Observation at the $x-$variable
j	Observation at the $y-$variable
b	Slope parameter
n	Total observation
N	Normal distribution
$f(x)$	Probability distribution of a function
s	Sample size
p	Number of parameters
q	Shape parameter
d	Specific observation
h	Height of a cluster tree
x	Observe value of x

y	Observe value of y
V	Residual value
P	Imputed values
O	Observed data values

LIST OF ABBREVIATIONS

BAB	Branch and Bound
COVRATIO	Covariance Ratio
DIFFITS	Difference in fits
DFBETA	Difference in Beta
EB	Estimated Bias
EIVM	Errors-in-variables model
EM	Expectation-maximization
EMB	Expectation-maximization with bootstrapping
LFRM	Linear Functional Relationship Model
LFRM1	Linear Functional Relationship Model when slope parameter is estimated using a MLE approach
LFRM2	Linear Functional Relationship Model when slope parameter is estimated using a nonparametric approach
LMS	Least Median of Squares
LTA	Least Trimmed Sum of Absolute Deviations
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MAR	Missing at Random
MCAR	Missing Completely at Random

MNAR	Missing Not at Random
MLE	Maximum Likelihood Estimation
MSE	Mean Square Error
pmask	Probability of Masking
pop	"Success" Probabability
pswamp	Probability of Swamping
SD	Standard Deviation
RMSE	Root-mean-square Error

LIST OF APPENDICES

CHAPTER 1: RESEARCH FRAMEWORK

1.1 Background of the Study

Errors-in-variables model (EIVM) or known as measurement error model has become an important topic since a century ago when studying the relationship between variables. It dates back in 1878 when Adcock wanted to fit a straight line to bivariate data when the bivariate information is measured with error. Since then, the EIVM study has been expanded and several literatures can be found over years (Lindley (1947), Madansky (1959), Anderson (1976), Fuller (1987), Gillard and Iles (2005), Tsai (2010)).

EIVM are regression models that take into account the measurement errors in the independent variables (Koul and Song, 2008). In contrast, the standard regression model assumes that the variables involved are measured exactly, or observed without error. If errors in the explanatory variables are ignored, the estimators obtained by classical or traditional regression are biased and inconsistent (Buonaccorsi, 1996). In real life, for example in biology, ecology, economics and environmental sciences, the variables involved cannot be recorded exactly (Gencay & Gradojevic (2011)).

To give an example, in the field of environmental sciences, measuring the level of household lead is an error-prone process as lead levels are exposed to many other media such as air, dust, and soil with possibly correlated errors (Carroll, 1998). Another example, when measuring nutrient intake, measurement error in a nutrient instrument can also be very huge, as there are daily and seasonal variability of an individual's diet thus resulting in the loss of power to detect nutrient-cancer relationship. In studies which include the case-control disease and serum hormone levels, measurement error also occurs due to a within-individual variation of hormones and also various laboratory errors. Therefore in real life examples, when the purpose is to estimate the relationship

between groups or populations, measurement errors arise (Patefeild (1985), Elfessi and Hoar (2001), Gillard (2007)).

Over the past 50 years, many researchers have been working on the problem of estimating the parameters in the linear functional relationship model (LFRM), a subtopic in the EIVM. However, the methods in the literature are mostly based on normality assumption, and it can be erroneous to use the normality assumption when there are outliers in the data set (Al-Nasser and Ebrahem, 2005). In other words, when there are outliers, a robust method is necessary to diminish the effect of the outlier. In 2005, Al-Nasser and Ebrahem proposed a new nonparametric method to estimate the slope parameter in a simple linear measurement error model in the presence of outliers. The nonparametric estimation method is a statistical inference which does not depend on a specific probability distribution. A significant advantage of using nonparametric method is that it is robust to outliers. This research has extended the study by Al-Nasser and Ebrahem (2005), by proposing a robust nonparametric method to estimate the slope parameter in LFRM.

Another area of the research is on identifying outliers, namely detecting a single outlier and multiple outliers in LFRM. An outlier is a point or some points of observation that is outside the usual standard pattern of the observations. Outlier occurs when the data is mistakenly observed, recorded, and inputted into the computer system (Cateni et.al., 2008). In linear models, Rahmatullah Imon (2005) and Nurunnabi et al. (2011) proposed group deleted version to identify outliers. In this study, the suitability of the *COVRATIO* procedure will be considered in detecting a single outlier for the data in the LFRM. The reason for choosing *COVRATIO* is that it is simple and is widely used in detecting outliers (Belsley et al., 1980). As mentioned earlier, the presence of multiple outliers situation are also taken into account. For multiple outliers, the clustering technique is considered, a method that is widely used to identify multiple outliers in a linear regression

model (Serbert et al., 1998; Adnan, 2003; Loureiro et al., 2004). In this study, the algorithm is developed that caters for data that can be model by the LFRM, where both the measurements are subject to errors.

The third area of this research is on the analysis of missing value in data sets. Missing data is unavoidable and is a significant problem that needs to be address. Some reasons that may cause the data to be missing include equipment malfunctioned, mistakes done during data entry, questions being omitted by respondents, and a subject being discarded due to the insufficient health condition. In this study, the two modern imputing approaches namely expectation-maximization (EM) and expectation-maximization with bootstrapping (EMB) are proposed for two kinds of LFRM models, namely LFRM1 for linear functional relationship model when slope parameter is estimated using a maximum likelihood estimation approach and LFRM2 for linear functional relationship model when slope parameter is estimated using a nonparametric approach.

1.2 Problem Statement

The area of parameter estimation in LFRM has been studied by several authors (Lindley, 1947; Kendall & Stuart, 1973; Wong, 1989; and Gillard & Illes, 2005). However, there has been insufficient work on the robust slope parameter estimator in LFRM.

In the first part of this study, the unidentifiable problem is overcomed by proposing a robust nonparametric method to estimate the slope parameter in LFRM. The second part of this study is related to the outlier problem and missing value problem in analysing quantitative data. It is crucial to identify a single outlier and multiple outliers as they give a tremendous impact in the statistical analysis stage. Several studies have been done on the identification of outliers problem in the linear regression model and circular regression model (Belsley et al., 1980; Rousseeuw & Leroy, 1987; Maronna et al., 2006, Ibrahim et al., 2013). However, methods of identifying outliers in the linear functional model are somewhat limited. Another common problem when analysing quantitative data is the presence of missing values (Little & Rubin, 1989). Missing data in the regression model and structural equation modeling (Little, 1992; Allison, 2003) has received a massive attention among researchers, however missing data in linear functional model has not received much attention. Therefore, in this study, the methods of handling missing data in LFRM is addressed.

1.3 Objectives of Research

The primary objective of this study is to propose a new robust parameter estimation and outlier detection method for linear functional relationship model. The specific objectives of this study are:

1. to propose a robust technique using nonparametric method to estimate the slope parameter in LFRM.
2. to propose the *COVRATIO* technique in detecting a single outlier in LFRM.
3. to propose the clustering technique in identifying multiple outliers in LFRM.
4. to identify a feasible modern imputation technique in handling missing values problem in LFRM.

Model verification of all the proposed method performed in this study is done by simulation studies. The applicability of the models is illustrated using Goran et al. (1996) data sets and two classical data used by Serbert et al. (1998).

1.4 Flow Chart of Study and Methodology

The flow chart of this study is outlined in Figure 1.1. First, a thorough literature review is conducted on the history and current issues and problems related to the errors-in-variable model, linear functional relationship model (LFRM), nonparametric estimation, outliers, and missing values. From the literature review, a robust method is developed using the nonparametric procedure for the slope parameter in LFRM. Then the robustness of this proposed method is compared with the existing Maximum Likelihood Estimation (MLE) method as well as with Al-Nasser and Ebrahem (2005) method.

Next, the *COVRATIO* technique to detect a single outlier for LFRM and propose a clustering technique to detect multiple outliers in LFRM is proposed. Finally, the missing values in LFRM is identified using the modern imputation technique. For the topics mentioned, simulation studies are conducted using S-Plus and R Programming to assess the performance of the proposed methods. The proposed methods are applied in real data sets for practical and illustration.

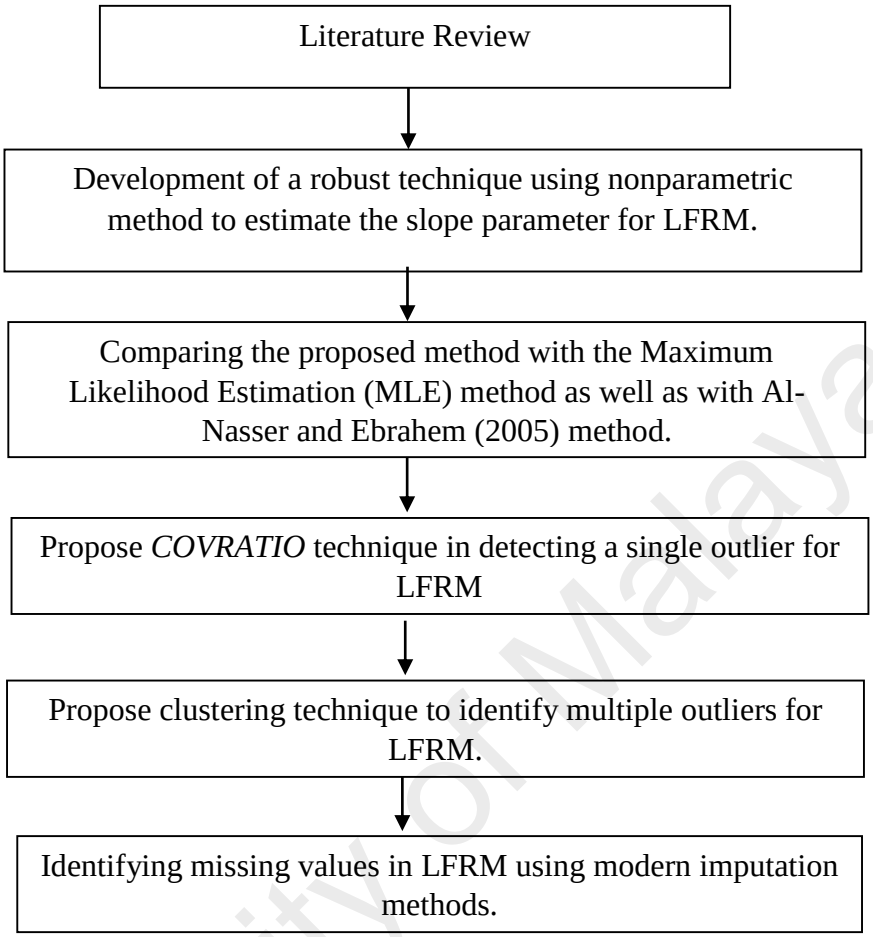

Figure 1.1: Flow chart of the study

1.5 Source of Data

In this study, the following data for illustration and application are used. Full data sets are given in Appendix A. The following are the background of the data sets used in this study.

1) **Goran et al. (1996) data**

   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).

2) **Hertzsprung-Russel Star Data**

   The data in Rousseeuw and Leroy (1987) are based on Humphreys et al. (1978) and Vansina and De Greve (1982) where 47 observations correspond to the 47 stars of the CYG OB1 cluster in the direction of Cygnus. The x variable in the second column is the logarithm of the effective temperature at the surface of the star, (T_e), and the y variable in column 3 is its light intensity (L / L_0). This data set contains four substantial leverage points which are the giant stars that corresponds to observations 11, 20, 30, and 34 that greatly affect the results of the regression line.

3) **Telephone Data**

   In this telephone data, Rousseeuw and Leroy (1987) give data on annual numbers of Belgian's phone calls, with x variable is the year from 1950 to year 1973, and y variable in the next column is the number of calls in tens of millions.

1.6 Thesis Organization

This thesis consists of seven chapters. Chapter 1 discusses the research framework which includes the background of EIVM, followed by the research objectives and the flow of the study. Chapter 2 reviews the literature and historical background of the research topics in this study. Chapter 3 proposes a robust nonparametric method to estimate the slope parameter in LFRM while Chapter 4 proposes a *COVRATIO* statistic to detect an outlier in the LFRM. Chapter 5 further extends the outlier problem by proposing the clustering technique to detect multiple outliers in LFRM. Chapter 6 reviews the missing value estimation methods for data that are in LFRM. Finally, Chapter 7 concludes the research findings and highlights some suggestion for future works.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter reviews the errors in variable model (EIVM) and the theoretical framework of the subtopic in EIVM, particularly the linear functional relationship model (LFRM). A brief historical review on the parameter estimation of LFRM is given. This section reviews the background information on the topics of outliers, particularly the single outlier detection method and the multiple outliers detection method. A literature review on the traditional and modern missing values problem is given at the end of this chapter.

2.2 Errors-in-Variable Model

Errors-in-variables model (EIVM) has been an important topic since a century ago, when Adcock (1878) investigated the estimation properties in ordinary linear regression models when both variables x and y are subject to errors with a restrictive but realistic assumptions. If the errors in the explanatory variables are ignored, then the estimators obtained using ordinary linear regression will be biased and inconsistent. Adcock obtained the least squares solution for the slope parameter by assuming both variables have equal error variance. In 1879, Kummel extended this study by assuming the error variance is known, but not necessarily equal to one. Later on in 1901, Pearson extended Adcock's findings of the equal error variance, to finding a solution for the $p-$ variate situation. Later on Deming's (1931) proposed orthogonal regression which was then included in his book and this method is sometimes known as Deming's (1931) regression.

In 1940, Wald proposed a different approach which does not take into account the error structure. Wald divided the order of the explanatory variables into two groups and

used the mean for the group to obtain the slope estimator. Later on, to get a more efficient estimator for the slope, Bartlett (1949) developed the grouping method by splitting the order of the explanatory variables into three groups, instead of two. Several grouping methods to group the explanatory variables has been reviewed by Neyman and Scott (1951), and Madansky (1959).

Another parameter estimation procedure that has been used in EIVM is the methods using the moments. Geary (1949) published an article using the method of moments. This is followed by Drion (1951) which uses the moments method and obtained new findings on the variance of the sample moments. Other studies on method of moments are by Pal (1980) and Van Montfort (1989) which focuses on getting optimal estimators using estimators that is based on higher moments.

Lindley and El-Sayyad (1968) proposed a Bayesian approach in EIVM regression problem and concluded that the likelihood approach may be misleading in some ways. Later on, Golub and Van Loan (1980) and Van Huffle and Vanderwalle (1991) introduced the total least square method in estimating the parameters in EIVM.

Application of EIVM can be shown in several fields. The total least square method has been widely used in dealing with optimization problem with an appropriate cost function in computational mathematics and engineering. Doganaksoy and van Meer (2015) have also applied the EIVM model in semiconductor device to assess their performance.

A new approach using the application of wavelet filtering approach which does not require instruments and gives unbiased estimates for the intercept and slope parameters has been introduced by Gencay and Gradojevic (2011). However, this approach still requires a lot more research, for example in cases with less persistent regressors. Another work by O'Driscoll and Ramirez (2011) focuses on the geometric view of EIVM. This method measures the errors using a geometric view to have an insight

on various slope estimators for the EIVM, which includes an adjusted fourth moment estimator proposed by Gillard and Iles (2005) in order to remove the jump discontinuity in the estimator of Copas (1972).

To summarize, the EIVM area of research has gain wide attention in studying the relationship between variables and dates back to as early as 1878.

To elaborate on the EIVM model, consider the following equation,

$$Y = \alpha + \beta X \,, \tag{2.1}$$

where both variables X and Y are linearly related but both are measured with error. Parameter α is the intercept, and β is the slope parameter. In reality, these two variables are not observed directly as their measurements are subject to error. For any fixed X_i, the x_i and y_i are observed from continuous linear variable subject to errors δ_i and ε_i respectively, i.e.

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

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

$$\delta_i \sim N\left(0, \sigma_\delta^2\right) \text{ and } \varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right). \tag{2.3}$$

This shows that the variances of error term are not dependent on i and therefore are independent of the level of X and Y. Substituting equation (2.3) into equation (2.2), the following equation is obtained,

$$y_i = \alpha + \beta x_i + \left(\varepsilon_i - \beta \delta_i\right). \tag{2.4}$$

This shows that the observable errors x_i and y_i are correlated with the error term $\left(\varepsilon_i - \beta \delta_i\right)$ and is independent of the slope parameter, β.

There are three models under the EIVM, namely the functional relationship, structural relationship, and ultrastructural relationship model as mentioned by Kendal and Stuart (1973), and are given as follows:

i) Functional relationship model between X and Y, is when X is a mathematical variable or fixed constant.

ii) Structural relationship model between X and Y, is when X is a random variable.

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

This study will focus on the linear functional relationship model (LFRM) which defines the X variable as a mathematical variable.

2.2.1 Linear Functional Relationship Model (LFRM)

As mentioned earlier, the linear functional relationship model (LFRM) is one example of an EIVM, which the underlying variables are deterministic (or fixed). Over the past three decades, many authors have been working on this functional model in EIVM (Lindley, 1947; Kendall & Stuart, 1973; Wong, 1989; and Gillard & Illes, 2005). Most of the study in LFRM have used maximum likelihood estimation method to estimate the parameters, with the assumption that the dependent and independent variables are joint normally and are identically distributed. Lindley (1947) first used the maximum likelihood estimation and realized that some assumptions on the parameter need to be made as there are some inconsistencies in the equation. Therefore, Lindley proposed the ratio of two errors to be known.

Since then, several authors did a rigorous research on handling the problem of estimating the parameters in LFRM. These findings include the geometric mean functional relationship by Dent (1935), two-group method of Wald and Wolfowitz (1940), maximum likelihood method by assuming known ratio of error variances by Lindley (1947), Housner and Brennan's method (1948), three-group method of Bartlett (1949), Durbin's ranking method (1954) and instrumental variables method mentioned by Kendall and Stuart (1961) and Fuller (1987). A detailed explanation for each method is given in Section 2.2.2.

Further study was done by Dorff and Gurland in 1961, and he extended this functional model as replicated and unreplicated functional relationship models, with certain recommendation. For unreplicated cases, the estimators by Wald and Wolfowitz (1940), Bartlett (1949) and Housner and Brennan's method (1948) have been considered and they found that Housner and Brennan's method (1948) of estimation is more robust than the Wald and Wolfowitz (1940) and Bartlett (1949) method and thus recommends the usage of it as compared to the others.

In the LFRM as given in equation (2.1) and (2.2), there are $(n+4)$ parameters, which are $\alpha, \beta, \sigma_\delta^2, \sigma_\varepsilon^2$, and the incidental parameters $X_1,...,X_n$. One complication arise as when the number of observations increase, the number of parameters will also increase. In this case when there is only a single observation at each point, the likelihood function is unbounded, and to overcome this problem, some constraint needs to be imposed, or the replicated data needs to be obtained. Some constraint includes making some assumptions on the variances and covariance of the errors, which includes:

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 $Cov(\delta_i, \varepsilon_i) = 0$.

Moberg and Sundberg (1978) mentioned that both the above conditions are necessary to find the maximum likelihood estimation of parameters in a linear functional relationship model with normally distributed errors. If only one of the error variances is known, then they show the likelihood equation for β is a cubic equation, which has a root corresponding to a plausible local maximum likelihood estimate of right sign only when the error variance is relatively small. This situation may cause the estimate to be inconsistent as the sample size increases. Another situation is to obtain replication of the information, which could be used to obtain consistent estimates of parameters, in particular for the β estimate. This research will focus on the estimate of β when replicates are not available.

In a linear functional relationship model, X and Y are mathematical variables which are linearly related, but are observed with error. For any fixed X_i, the x_i and y_i are observed from continuous linear variable, subjected to errors δ_i and ε_i respectively, i.e.

$$x_i = X_i + \delta_i \text{ and } y_i = Y_i + \varepsilon_i, \text{ where } Y_i = \alpha + \beta X_i,$$

$$\text{for } i = 1, 2, ..., n, \tag{2.5}$$

where the α is a constant and β is the slope function. The δ_i and ε_i are assumed to be mutually independent and normally distributed random variables, that is $\delta_i \sim N\left(0, \sigma_\delta^2\right)$ and $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$. This model as in (2.5) is known as the unreplicated linear functional relationship model as there is only a single observation for each level of i.

There are $(n+4)$ parameters to be estimated, which are $\alpha, \beta, \sigma_\delta^2, \sigma_\varepsilon^2$, and the incidental parameters $X_1, ..., X_n$. In estimating the parameters, the majority attention usually focuses on estimating β, that is the slope parameter, as from a theoretical viewpoint, the role of α, the intercept parameter is minor (Cai and Hall, 2006).

The log likelihood function is given by

$$\log L\left(\alpha,\beta,\sigma_\delta^2,\sigma_\varepsilon^2,X_i,...,X_n;x_1,...,x_n,y_1,...,y_n\right)=$$

$$-n\log(2\pi)-\frac{n}{2}\log\sigma_\delta^2-\frac{n}{2}\log\sigma_\varepsilon^2-\frac{\sum\left(x_i-X_i\right)}{2\sigma_\delta^2}-\frac{\sum\left(y_i-\alpha-\beta X_i\right)^2}{2\sigma_\varepsilon^2}. \qquad (2.6)$$

The likelihood in equation (2.6) is unbounded, let say when putting $\hat{X}_i = x_i$ and considering σ_δ^2 approaches to 0, the likelihood function will approach infinity, irrespective of the values of α,β and σ_ε^2. Therefore, to avoid an unbounded problem in this equation, additional constraint is assumed, $\sigma_\varepsilon^2=\lambda\sigma_\delta^2$, where λ is known (Lindley, 1947). The log likelihood function becomes

$$\log L\left(\alpha,\beta,\sigma_\delta^2,X_i,...,X_n;\lambda,x_1,...,x_n,y_1,...,y_n\right)=$$

$$-n\log(2\pi)-\frac{n}{2}\log\lambda-n\log\sigma_\delta^2-\frac{1}{2\sigma_\delta^2}\left\{\sum\left(x_i-X_i\right)^2+\frac{1}{\lambda}\sum\left(y_i-\alpha-\beta X_i\right)^2\right\}. \qquad (2.7)$$

There are $(n+3)$ parameters to be estimated, namely $\alpha,\beta,\sigma_\delta^2$ and the incidental parameters, $X_1,...,X_n$. Differentiating $\log L$ with respect to parameters $\alpha,\beta,\sigma_\delta^2$ and X_i, the parameters $\hat{\alpha},\hat{\beta},\hat{\sigma}_\delta^2$ and $\hat{X}_i$ can be obtained, given by

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

$$\hat{\beta}=\frac{S_{yy}-\lambda S_{xx}+\left\{\left(S_{yy}-\lambda S_{xx}\right)^2+4\lambda S_{xy}^2\right\}^{\frac{1}{2}}}{2S_{xy}},$$

$$\hat{\sigma}_\delta^2=\frac{1}{(n-2)}\left\{\sum\left(x_i-\hat{X}_i\right)^2+\frac{1}{\lambda}\sum\left(y_i-\hat{\alpha}-\hat{\beta}\,\hat{X}_i\right)^2\right\},$$

and

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

where $\bar{y} = \frac{1}{n}\sum y_i$, $\bar{x} = \frac{1}{n}\sum x_i$,

$$S_{xx} = \sum (x_i - \bar{x})^2,\ S_{yy} = \sum (y_i - \bar{y})^2 \text{ and } S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) . \qquad (2.8)$$

Further details of the parameter estimation can be found in the literature (Sprent 1969, Kendall and Stuart 1973, Al-Nasser and Ebrahem, 2005). As for the variance of the parameter estimate, Patefield in 1977 derived a consistent asymptotic covariance matrix of the ML estimates for α and β by partitioning the following information matrix, given by

$$\begin{pmatrix} \hat{Var}(\hat{\alpha}) & \hat{Cov}(\hat{\alpha}, \hat{\beta}) \\ \hat{Cov}(\hat{\alpha}, \hat{\beta}) & \hat{Var}(\hat{\beta}) \end{pmatrix},$$

where $\hat{Var}(\hat{\alpha}) = \frac{(\lambda + \hat{\beta}^2)\,\hat{\sigma}_\delta^2 \hat{\beta}}{S_{xy}} \left\{ \bar{x}^2 (1 + \hat{T}) + \frac{S_{xy}}{n\hat{\beta}} \right\}$,

$$\hat{Var}(\hat{\beta}) = \frac{(\lambda + \hat{\beta}^2)\hat{\sigma}_\delta^2 \hat{\beta}}{S_{xy}} \{1 + \hat{T}\}, \text{ and}$$

$$\hat{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{(\lambda + \hat{\beta}^2)\hat{\sigma}_\delta^2 \hat{\beta}\,\bar{x}}{S_{xy}} \{1 + \hat{T}\},$$

where , $\hat{T} = \frac{n\lambda\hat{\beta}\hat{\sigma}_\delta^2}{(\lambda + \hat{\beta}^2) S_{xy}}$. $\qquad (2.9)$

2.2.2 Parameter Estimation of Linear Functional Relationship Model

As mentioned in Section 2.2.1, one complication arises in LFRM, as when the number of observations increase, the number of parameters will also increase. When the LFRM has only a single observation at each point, the likelihood function is unbounded, and to overcome this problem, some constraint is imposed or the replicated data is obtained. As mentioned, Lindley (1947) propose the case when the ratio of the error variance λ is known. This study focuses on the slope parameter estimation for LFRM as knowledge on the slope parameter is also crucial.

From literature, there are several methods of estimating the slope parameters. Dent in 1935 propose the geometric mean functional relationship estimator, which is

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

and this slope estimator has been widely used in fisheries research. This estimator is symmetric in both x and y and thus still preserve the inherent symmetry of the functional relationship model. Sprent (1969) mentioned that this estimator has an intuitive appeal, but is usually not consistent, as it only ignores the identifiability problem, and assumes normality without knowing the error variance.

Later on Wald (1940) proposed a two-group method to find a consistent estimator for β. He computed the arithmetic means $\left(\bar{x}_1, \bar{y}_1\right)$ for lower group of observations. Then the higher group of observations, $\left(\bar{x}_2, \bar{y}_2\right)$ is computed, after it is arranged in ascending order by the basis value of x_i. Then, these values are divided into two equal sub-groups, and the slope parameter is estimated by,

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

This estimation method gives consistent estimate of β, even though it is not the most efficient as its variance does not have the smallest possible values. However, it seems that this method of estimation is not symmetric in x and y, as the upper and lower groups are not necessarily the same when ranked on y_i. One way to make this method symmetric is by taking the average of this with the equivalent one based on ranking them by the base of the y_i.

Next, in 1949 Bartlxett proposed the method which is same idea with the two-group method, that is the observations are arranged in ascending order on the basis of x_i values, and he extended the method by dividing them into three equal groups. If the number of observations is not exactly divisible by 3, then he will make it approximately equal. The middle group will be ignored, then the arithmetic means $(\bar{x}_1, \bar{y}_1)$ for the lowest group and $(\bar{x}_3, \bar{y}_3)$ for the highest group is calculated, and the slope parameter β is estimated using this formula,

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

This method generally gives a consistent estimate for β, and performs more efficient than the two-group method. However, the estimator is not symmetric in x and y, as the upper and lower groups are not necessarily the same when ranked on base on y_i.

Housner-Brennan (1948) proposed a consistent estimate of β, where first, the x_i values are arranged in ascending order, as $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, and the associated values of y which may not be in ascending order are taken. The estimate of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} i\left(y_{(i)} - \bar{y}\right)}{\sum_{i=1}^{n} i\left(x_{(i)} - \bar{x}\right)}, \tag{2.13}$$

however, this slope estimator is not symmetric in x and y.

Durbin's "ranking" method (1954), suggested that the estimate of β is given by,

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

where x's and y's are ranked in ascending order, on the basis of x values. Later on interchange them and arrange the y values in ascending order. From this proposed method, the estimator is still not symmetric in x and y.

Cheng and Van-Ness (1999) then proposed the modified least squares, when the variance ratio of $\lambda = \frac{\tau^2}{\sigma^2}$ is assumed to be known. The slope estimator will be,

$$\hat{\beta} = \frac{\left(S_{yy} - \lambda S_{xx}\right) + \left(\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S^2{}_{xy}\right)^{\frac{1}{2}}}{2S_{xy}}, \tag{2.15}$$

where

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

The method proposed here leads to the same estimates as mention in Section 2.2.1, but without requiring the normality assumption.

Al-Nasser and Ebrahem in 2005 proposed a nonparametric approach for the slope parameter, where it does not require a normality assumption. A nonparametric procedure has several strengths, such as no prior knowledge on the distribution of the model is needed, and in the presence of "noises" in a data set, this nonparametric procedure will still be useful to estimate the trends of the data (Sprent & Smeeton, 2016). In his proposed method, the x_i values are arranged in ascending order, as $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ and the associated values of y which may not be in ascending order are taken. He then listed down all the possible paired of slopes and find the median of all the slopes listed to be the final slope parameter.

From the above literature, only few studies use nonparametric assumption. Al-Nasser and Ebrahem (2005) studied on the parameter estimation method when outliers are present in the data. However, this method is only robust when the outliers is 20% or more of the total observation. It is also crucial to identify outliers as low as 1%, 5% and 10% from the total observation. In this research, a robust nonparametric estimation method which is an extension from the study by Al-Nasser and Ebrahem (2005) method in the presence of outliers is proposed and will be elaborated in Chapter 3.

2.3 Outliers

In this section, the observation that gives a huge impact in data analysis namely the outliers are discussed. The study of outliers is very important and is considered to be as old as the subject of statistics. An outlier is a point or some points of observation that is outside the usual pattern of the other observations. As mentioned by Chen et al. (2002) "Outliers are those data records that do not follow any pattern in an application". Outlier occurs when the data is mistakenly observed, recorded, and inputted in the computer system (Cateni, 2008). According to Hampel et al. (1986), it is common to have 1% to 10% of outliers in a data set; in fact, the data set that has the best quality is also prone to have at least a very small amount of outliers. Studies on outliers in linear model can be seen in Wong (1989), Cheng and Van Ness (1994) and Elfessi and Hoar (2001), Satman (2013), and Hussin et al. (2013).

In fitting a linear regression model by the least squares method it is often observed that a variety of estimates can be substantially affected by one observation or a few observations (Rousseeuw and Leroy (1987), Maronna et al. (2006)). It is important to locate such observations and assess their impact on the model, either it gives a huge impact to the model or just a low impact on the model.

An outlier is a point that falls away from the other data points. If the parameter estimates change significantly when a point is removed from the calculation, then this point is considered to be influential. From Figure 2.1, one outlier can be seen. This outlier lies away from the other observations. When including outlier 1 in the analysis of the least square regression and plotting the points, the black line is produced. However, if the outlier is deleted, a new regression line is obtained, which is the red line. This means that outlier 1 is an influential observation, as it changes the regression line and there is an extreme value in Y.

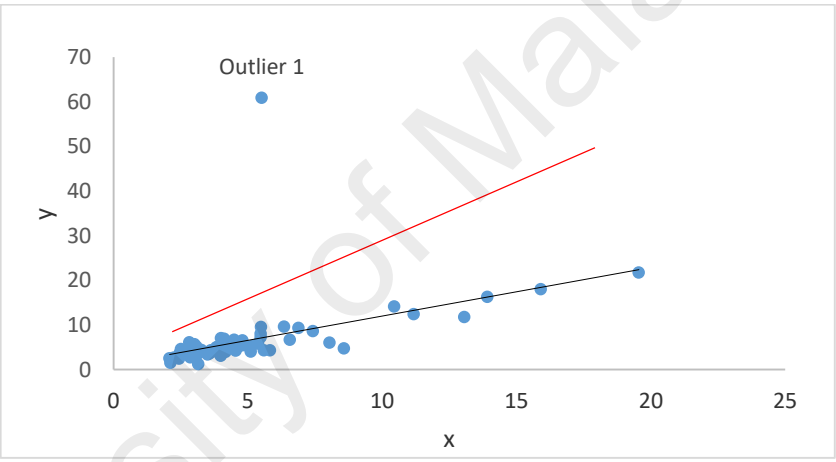

Figure 2.1: Example of an outlier

Next, the leverage point. Points with extreme values of X are said to have high leverage, which means that high leverage points have a greater ability to move the line. As an example, outlier 2 in Figure 2.2 is a high leverage point, because when removing this outlier, the regression line shifts from the black line to the red line. Outlier 3 on the other hand, is a good leverage as when removing this point, it does not change the regression line.

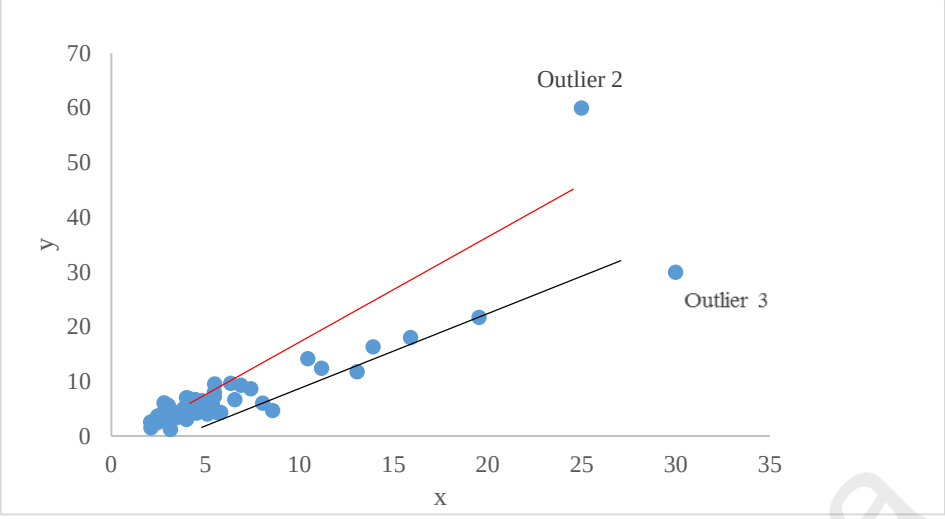

Figure 2.2: Example of a high leverage X point.

A number of outlier diagnostics are available in the literature include Cook's distance, Difference in fits (DIFFITS), Difference in Beta (DFBETA), Covariance Ratio (*COVRATIO*) (Belsley et al., 1980) and many others.

Cook (1979) proposed a measure of Cook's Distance, CD_i using the studentized residuals and the variances of residuals and predicted values. The *i*th Cook's distance provides a measure of how much the parameter estimates change when a point is remove from the calculation, which is introduced as

$$CD_i = \frac{\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)^T \left(X^T X\right)\left(\hat{\beta}^{(-i)} - \hat{\beta}\right)}{k\hat{\sigma}^2}, \tag{2.16}$$

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

The *i*th difference in fits (DFFITS) is also used to show how influential a point is in a statistical regression, and is defined by

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}, \quad i = 1, 2, ..., n \tag{2.17}$$

where $\hat{y}_i^{(-i)}$ are the fitted responds, $\hat{\sigma}_{(i)}$ are the estimated standard error when the *i*th observation is deleted and h_{ii} is the leverage. A small value of *DFFITS* indicates a low leverage point.

DFBETAS statistics are used to measure the change in each parameter estimate and are calculated by deleting the i^{th} observation,

$$DFBETAS_j = \frac{b_j - b_{(i)j}}{s_{(i)}\sqrt{\left(X^{'}X\right)_{jj}}}, \tag{2.18}$$

where $\left(X'X\right)_{jj}$ is the $(j,j)^{th}$ element of $\left(X^{'}X\right)^{-1}$. A large value of *DFBETAS* indicate that the observations are influential in estimating the parameter.

Another measure of outliers is *COVRATIO* which is use as a statistical measure to identify the change in the determinant of the covariance matrix of the estimates by deleting the i^{th} observation, and is defined by

$$COVRATIO_{(-i)} = \frac{\left|COV\right|}{\left|COV_{(-i)}\right|}, \tag{2.19}$$

where $\left|COV\right|$ is the determinant of covariance matrix of full data set and $\left|COV_{(-1)}\right|$ is that of the reduced data set by excluding the i^{th} row. *COVRATIO* has been well established in regression modelling by Belsley et. al. (1980) and has also been used in functional relationship model for circular variable by Hussin and Abuzaid (2012). Recently, Ibrahim et al. (2013) identified outliers in circular regression model by using the *COVRATIO* procedure. In LFRM, however, methods of identifying outliers are somewhat limited. As this simple linear functional relationship model has a close resemblance of the linear regression model, and due to its simplicity and widely usage, the *COVRATIO* technique in detecting a single outlier will be proposed in this LFRM in Chapter 3.

2.3.1 Cluster Analysis

Outlier cases happen when there is a single outlier or when there are multiple outliers. Identifying a single outlier is quite simple from the analytical and computational side, but when there is more than one outlier, then it becomes even challenging. Identifying multiple outliers become more complicated due to masking and swamping effects. Masking happens when an outlier is unable to be detected as a true outlier, while swamping happens when a "clean" observation, or an inlier is falsely detected as an outlier. Masking seems to be a more serious issue than swamping, but both these effects should be identified so that appropriate analysis can be done on the data set (Sebert et al., 1998).

In general, there are two ways to classify the multiple outlier detection procedures, which are the direct method and the indirect method (Hadi and Simonoff, 1993). The direct method are procedures base on least square and are specifically designed algorithm to detect multiple outliers. The indirect method on the other hand, uses the result from robust regression estimates, and when there are outliers, the least square methods will differ significantly from when there is no outlier.

Some direct methods include the study by Swallow and Kianifard (1996). In this study, they suggest that recursive residuals to be standardized by a robust estimate of scale, to classify the multiple outliers. Sebert et al. (1998) proposed a clustering algorithm using the single linkage algorithm and Euclidean distance, which helps to find the single largest cluster, and identify them as inliers. Fernhloz et al. (2004) proposed a new method for detecting outliers based on the multihalver, or known as the delete-half jacknife and is also applicable for multivariate data.

The indirect method is through a robust regression estimate, which includes the techniques by Rousseeuw (1984), Hawkins and Olive (1999) and Agullo (2001). Rousseeuw (1984) introduced the high breakdown (as high as 50%) for Least Median of

Squares (LMS) estimator whereby the LMS estimator $\hat{\beta}$ is obtained from minimizing the median of squared errors. Hawkins and Olive (1999) proposed the use of least trimmed sum of absolute deviations (LTA) as an alternative to LMS, where the computational complexity is lower than the LMS. The LTA is particularly attractive for large data sets and it is used as a tool for modelling data sets that deals with missing values on the predictors. In 2001, Agullo proposed two new algorithms to compute the LTS estimator, where the first algorithm is probabilistic and refer to the exchange procedure. The second algorithm is exact and is based on a branch and bound (BAB) technique that guarantees global optimality and without exhaustive evaluation. The BAB is computationally feasible for $n \leq 50$ and $p \leq 5$, which seems to be a very small data set.

In this study, the focus will be on the direct method to identify multiple outliers, namely the clustering procedure. Several studies have been using clustering procedure for the outliers problem, such as detecting outliers in regression model (Sebert et al., 1998; Adnan and Mohamad, 2003), and detecting erroneous data in foreign trade transaction (Loreiroe et al. 2004). However, detecting outliers using clustering method has not been explored for LFRM.

As the linear regression model resembles the LFRM, the clustering algorithm as proposed by Sebert et al. (1998) to identify multiple outliers will be developed for this LFRM. Sebert et al. (1998) cluster analysis begins by taking a set of n observations on p variables. Next, a measure of similarity between observations are obtained, by employing a certain inter-observation similarities. An important procedure that one must decide before applying the clustering algorithm is the variables to use, the measure of similarity to use, and finally which clustering algorithm to use.

2.3.2 Similarity Measure for LFRM

To group the "variables" or items into their own groups, it is necessary to have a certain measurement of "similarity" or a measure of dissimilarity between the items. There are four types of similarity measure which are correlation coefficient, distances measures, association coefficients and probabilistic similarity coefficients (Aldenderfer & Blashfield, 1984).

All these four methods have its own strengths and drawbacks, so it is necessary to choose the best measurement that fits the model. The most commonly used similarity measure is Euclidean distance, defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2} , \tag{2.20}$$

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

Another type of measurement distance or known as the city-block metric is the Manhattan distance, which is defined by

$$d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{ij} \right|^r \right)^{\frac{1}{r}} . \tag{2.21}$$

Minkowski metrics which is a more specific forms of the special class of metric distance function can be defined as

$$d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{ij} \right|^r \right)^{\frac{1}{r}} . \tag{2.22}$$

Another distance is the generalized distance (Malahanobis) which is defined as

$$d_{ij} = \left(X_i - X_j \right) \sum {}^{-1} \left(X_i - X_j \right) \tag{2.23}$$

where $\sum$ 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.

For this LFRM model, the Euclidean distance will be used as the similarity measure. Euclidean distance has been widely used and commonly accepted when grouping multivariate observations (Everitt, 1993). Euclidean distance, defined as in equation (2.20) has been popular because it can be easily applied, where by similar observations are identified by relatively small distance, while a dissimilar observation is identified by a relatively large distance.

2.3.3 Agglomerative Hierarchical Clustering Method

As mentioned by Estivil-Castro (2002), it is important to understand the "cluster model" as this is the key to differentiate each of these clustering algorithm. The typical cluster model includes the following. First is the connectivity models as an example, the hierarchical clustering builds models which is based on distance connectivity. Next, the centroids models for example, the k-means which represents each cluster by its mean. The distribution models on the other hand, clusters the observation using a statistical distribution. Another cluster model is the density model that defines clusters as connected dense regions in a certain data space. Besides that, a group models cluster the observation by just providing the grouping information. And finally, a graph-based model which is a subset of nodes in a graph where every two nodes in the subset are connected by an edge can be identified as a form of cluster. Each of these models represent a different algorithm and it is important to choose a specific clustering method that is compatible with the nature of the classification in this field of study.

Among the most popular used algorithm is the hierarchical clustering as it is simple and easy to use (Dasgupta and Long, 2005). This type of cluster is useful for analyst as it requires no prior specification of the number of clusters. This hierarchical cluster operates based on the similarity matrix in order to construct a tree depicting specified relationship between each observation. Figure 2.3 illustrates the branches and

root in a hierarchical clustering, where the agglomerative methods build a tree from branches to root, while the divisive methods build a tree from the root, and finishes at the branches.

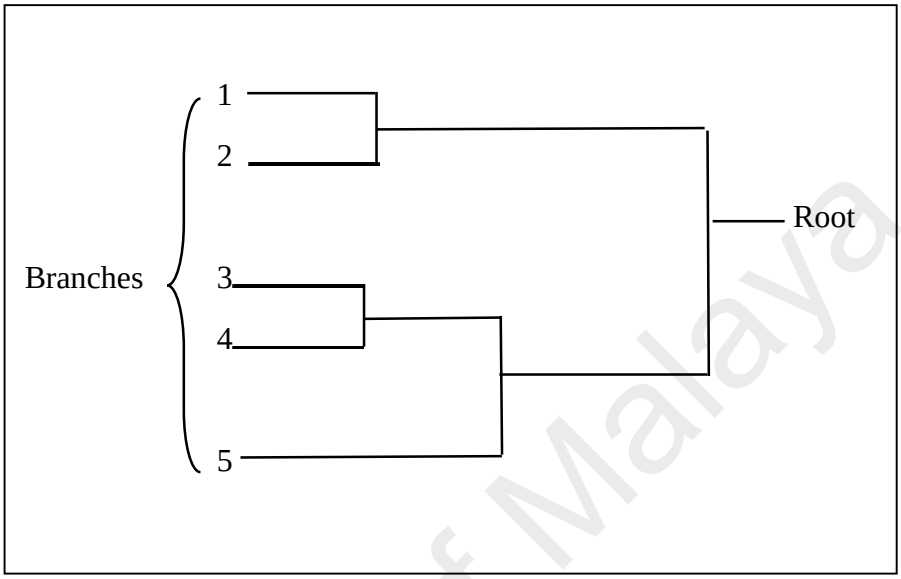

Figure 2.3: Illustration of branches and root in a hierarchical clustering methods.

The agglomerative hierarchical method begins with a series of successive merging between individual observations as clusters. First, the objects that have a similarity are grouped, then later on they are merged based on the similarity measure. As the similarity decreases, all the subgroups are fused in a single cluster and are nested, which means they are permanently merged together. The divisive hierarchical methods are the opposite of agglomerative, which means it builds a tree from the root, and finishes at the branches. 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.

There are three major clustering techniques in agglomerative hierarchical clustering as follows (Kaufman and Rousseeuw, 1990).

1. Linkage method
    - Single linkage (nearest neighbor), uses the smallest dissimilarity between a point in the first cluster and a point in the second cluster.
    - Complete linkage (farthest neighbor), uses the largest dissimilarity between a point in the first cluster and a point in the second cluster.
    - Average linkage (average neighbor), uses the average of the dissimilarities between the points in one cluster and the points in the other cluster.
2. Centroid methods use the Euclidean distances as the dissimilarity between two means of the clusters. The centre will move as the clusters are merged.
3. Ward's method or known as error sum of squares method. This method is basically looking at the analysis of variance problem, instead of using distance metrics or measures of association.

Representation of the major clustering techniques in agglomerative hierarchical are shown in Figure 2.4, where it can be seen that the single and complete linkage methods are simple (Mirkin 1998). Single linkage clusters are isolated and have a noncohesive shape, while the complete linkage clusters are very cohesive but is not isolated (Chowdury, 2010). The other linkages, namely the average, centroid and Ward method represent the "middle way" and are rather close to each other in order to construct a tree diagram (Mirkin 1998). Among the ways to cluster the data, single linkage is found to be

the easiest mathematically in constructing the clusters and has been widely used since it was introduced by Sneath and Sokal (1973) in the field of biology and ecology, and later on by Aldenderfer and Blashfield (1984) in computational statistics.

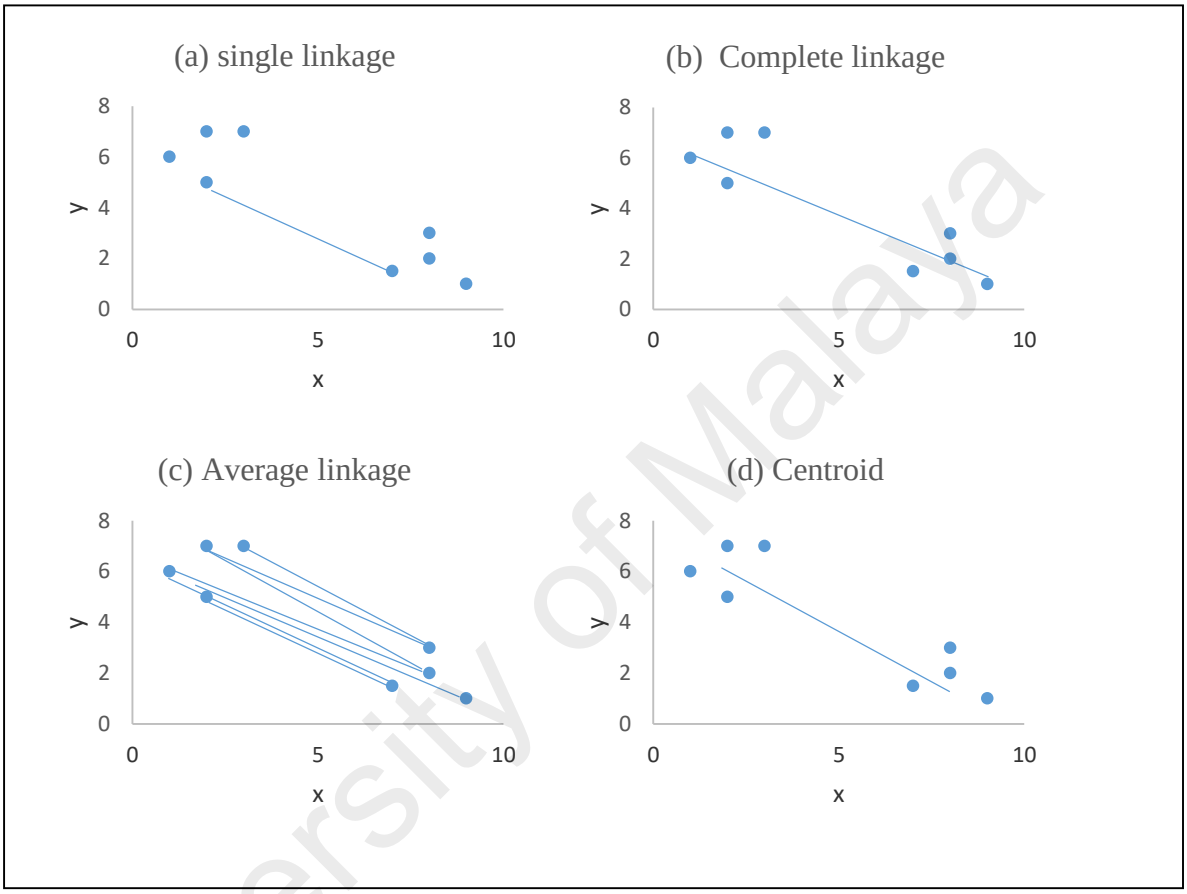

Figure 2.4: Representation of the major clustering techniques in agglomerative hierarchical; (a) Single linkage, (b) Complete linkage, (c) Average linkage, (d) Centroid

The focus of this study is on the single linkage method, as it is easy to compute, and as the area of multiple outliers in LFRM is new, a computationally easy approach is practically needed. Single linkage method operates on a similarity coefficient between groups, which is revised as each successive level of the hierarchical is generated. The

term single is used, because clusters are joined when the objects in different clusters have sufficiently small distances, as if a single link is use to connect the clusters. The inputs to this linkage is either the distances or similarities between pairs of objects. Then, the groups are formed from individual entities by merging nearest neighbours which is obtained from the smallest distance or from the entities with the largest similarities. This study attempts to develop a single linkage clustering algorithm technique for identifying multiple outliers in linear functional relationship model. A detail discussion on this topic is given in Chapter 5.

2.4 Missing Values Problem

Presence of missing value is unavoidable in all fields of quantitative research. They can be seen in the field of economics (Takahashi & Ito, 2013), medical (Dziura et al. 2013), environmental (Razak et al. 2014; Zainuri et al. 2015), life sciences (George et al. 2015), and social sciences (Acock 2005; Schafer & Graham 2002). It has been established that ignoring missing values may result in biased estimates and invalid conclusions (Little & Rubin, 1987; Guan & Yusoff 2011). There are several reasons that may cause a data to be missing. First is when nonresponse occur, where the item seems sensitive to individuals, thus they choose to leave the item blank, let's say the monthly income. Dropout may occur mostly when studying a research over a certain period of time, where a few participants may drop out before the experiment ends. Another reason why data may be missing is due to equipment malfunction or mistakes during data entry.

In the field of psychology, it is a real challenge for longitudinal research as the data obtain from a multiple wave of measurement on the same individual may cause it to be incomplete. From among 100 longitudinal studies obtained from three developmental journals- Child Development, Developmental Psychology, and Journal of Research on

Adolescence, 57 of the cases have been reported either having missing values or had discrepancies in sample sizes (Jelicic et al., 2009).

Impact of missing data is also a challenge in the field of gene expressions, where the experiments often contain missing values, due to insufficient resolution, image corruption, and due to contaminants such as dust or scratches on the chip (de Souto et al., 2015). In environmental research, obtaining the air quality data it will also be of a challenge as data are likely to be missing due to machine failure and insufficient sampling (Zainuri et al., 2015). In short, inadequate approach of handling missing data in a statistical analysis will lead to erroneous estimates and incorrect inferences.

Missing data can be classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). MCAR is when the missing in X variable is not related to any other variables, or the X variable itself. An example of MCAR situation is when a participant misses a scheduled survey, due to a doctor's appointment and not because of the things related to the survey question. Next, MAR mechanism is when the missing data is correlated with the other study-related variables in the analysis. As an example, the increase of substance usage, will relate to chronic absenteeism, leading to an increase in the probability of data missing for the self-esteem measure. The MNAR on the other hand is when the probability of missing data is completely related to the values that are missing. An example is when there are missing data on the reading scores and this is completely related to a person's reading ability (Baraldi & Enders, 2010).

In general terms, techniques to deal with missing values can be categorised as traditional or modern approach. Some review on the traditional and modern missing data techniques are given in the next section.

2.4.1 Traditional Missing Data Techniques

Some commonly used traditional ways are listwise deletion and pairwise deletion. As for imputation methods, mean imputation, hot-deck imputation, and stochastic imputation are among the commonly used ones (George et al., 2015). Listwise deletion is when an individual in a data set is deleted from an analysis if there are missing data on any of the variable in the study. It is a simple approach to handle the missing values and it gives a complete set of data, but it creates even larger problem to the statistical analysis stage. When the missing data are deleted, it reduces the sample size, and this is a huge disadvantage if the total number of missing item is high. Hence, lack of statistically significant estimates of conclusion occur (Tsikriktsis, 2005)

Another commonly used method in handling missing data is pairwise deletion or also known as the available case analysis (Peugh and Enders, 2004). In pairwise deletion, the missing data are removed on an analysis-by-analysis basis, such that when a particular variable has a missing value, other variables that has no missing values can still be used during the analysing stage. The pairwise deletion maximizes all the data that is available, thus increases the power in the analysis. However, the disadvantage of this pairwise deletion is that the standard of errors computed by most of the software packages uses the average sample size across analyses, thus making the standard of errors underestimated or overestimated.

Another common technique that is use in handling missing data is the single imputation method, which means the researchers imputes the missing data with some suitable replacement values (Baraldi and Enders, 2010). There are different types of imputation techniques, but the most common approach from the single imputation is mean imputation, regression imputation, hot-deck imputation and stochastic imputation. For mean imputation, the mean is obtained from the arithmetic mean of the available data are replaced in the missing values (Tsikriktsis, 2005; Baraldi and Enders, 2010). The mean

imputation is easy to use, but the variability in the data is reduced, thus making the standard deviation and variance estimates being underestimated. For example, if there is 20% of the total observation to be missing, this means that 20% of the data will have zero variance after the mean imputation technique is implied. This situation will be problematic, especially when the missing value is high (Acock, 2005).

In regression imputation, it involves a regression equation which uses the available data that are not missing, to predict the expected values for the missing data. To simplify, the missing values are the outcome variable, while the other variables in the data set are the predictors. The technique give a good "guess" as it obtains information from the complete variables, however this method produces biases in the variances and covariances (Graham et al., 2003).

Another imputation method is the hot-deck imputation method where this method is based on matching the case of the missing data with similar cases without missing data. Correlation matrix is used to determine the most highly correlated variables. It is an advantage if the sample is large, as similar case can easily be identified. However, a drawback of this method is that it involves a single value which reduces the amount of variations in the data (Schlomer et al., 2010). Stochastic regression implies the regression imputation with some modification, where a random value is added to the imputed predicted value. This imputation technique are centered at zero, therefore they do not change the mean, thus provide the same unbiased means as does the regression technique. However, with this stochastic values, it introduces variance in the imputed data, which results in unbiased variance estimates (Little & Rubin, 1987). These traditional methods are basically easy to use, but they also cause several drawbacks which makes these methods unlikely to be used by researchers. Some modern techniques to handle missing data has been developed few years later.

2.4.2 Modern Missing Data Techniques

Modern missing techniques are later introduced, which some of them are integrated from traditional techniques. These modern approaches include those based on maximum likelihood and multiple imputations (Acock 2005). Expectation-maximization (EM) algorithm is an example of maximum likelihood and some examples of multiple imputations include Markov Chain Monte Carlo, Fully Conditional Specification, and Expectation-maximization with bootsrapping (EMB) algorithm (Baraldi & Enders, 2010; Barzi & Woodward, 2004; Gold & Bentler, 2000; Little & Rubin, 1987).

Maximum likelihood estimation approach uses the available data, either complete or incomplete, and finds the parameter that will have the highest probability of creating the sample data. Multiple imputation approaches creates multiple copies of the data set and each copy will have different imputed values. Therefore for this approach, there will be three stages which is the imputation stage, analysis stage, and pooling results stage (Baraldi & Enders, 2010).

These modern approaches are superior to the traditional approach as they produce unbiased estimates for both MCAR and MAR data, and no data are "thrown out", thus making the method more reliable. In this research, a feasible modern imputation method is identified to apply in the data that can be modelled by the LFRM and will explain in detail in Chapter 6.

CHAPTER 3: NONPARAMETRIC ESTIMATION FOR SLOPE OF LINEAR FUNCTIONAL RELATIONSHIP MODEL

3.1 Introduction

In this chapter, a robust nonparametric method to estimate the slope parameter of a LFRM in which both parameters are subject to error is proposed. In this case, the error variance ratio is assumed known, that it is equal to one. Section 3.2 briefly describes the nonparametric method for estimating the parameter. Section 3.3 proposes a new estimation method for a slope parameter of the LFRM. A simulation study is conducted in Section 3.4 to compare the proposed slope estimation method with the existing MLE method and the nonparametric method by Al-Nasser and Ebrahem (2005). Simulation results and discussion are presented in Section 3.5. Finally, the application of a real life data using the proposed nonparametric estimation method is illustrated in Section 3.6.

3.2 Nonparametric Estimation Method of LFRM

As mentioned earlier for LFRM in Section 2.2.1, where $Y = \alpha + \beta X$, both of the two variables X and Y are observed with errors, with α and β is the intercept and slope parameter respectively. For any fixed X_i, the x_i and y_i are observed from continuous linear variable subject to errors δ_i and ε_i respectively. The error terms δ_i and ε_i are assumed to be mutually independent and normally distributed random variables, with mean 0 and variance σ_δ^2 and σ_ε^2 respectively. In this parametric LFRM, there are $(n+4)$ parameters that need to be estimated, namely $\alpha, \beta, \sigma_\delta^2, \sigma_\varepsilon^2$ and the incidental parameters $X_1, ..., X_n$. However, with these incidental parameters, it leads to inconsistencies of the estimators. Thus, some information is needed to overcome the

inconsistencies of the estimators, which is, either one of the variances or the ratio of the two variances is known (Fuller, 1987).

Several methods of estimation of LFRM have been suggested in previous studies such as Kendall and Stuart (1979), Fuller (1987), Cheng and Van Ness (1999), Huwang and Yang (2000) and Al-Nasser (2004). However, the methods in the literature are mostly based on normality assumption, and it can be erroneous to use the normality assumption when there are outliers in the data set. In other words, when there are outliers, a robust method is necessary to diminish the effect of the outlier.

The nonparametric estimation method is a statistical inference which does not depend on a specific probability distribution. As mentioned by Hajek (1969), this method is widely used and also easy to perform. An important advantage of using nonparametric method is that it is generally robust to outliers. A number of researchers have studied the nonparametric estimation methods, such as Dent (1935), Housner-Bernnan's (1948), Theil (1950) and Cheng and Van-Ness (1999). These traditional estimation methods however, do not consider the presence of outliers in the data. A study by Al-Nasser and Ebrahem (2005) which incorporated the presence of outliers, compared his proposed method with the other traditional estimators. From the study, when the percentage of outliers are 20% or more, Al-Nasser and Ebrahem (2005) method seems to be robust to outliers, unlike all the other traditional methods. However, in real life experiment, outliers also exist in small amount of numbers. Hence, it is crucial to identify when there are small percentage of outliers in an experiment, such as when there is a single, 5% and 10% of outliers. Therefore, in this chapter, a robust nonparametric method to estimate the slope parameter in LFRM is proposed by further improving the nonparametric method as proposed by Al-Nasser and Ebrahem (2005), and comparing it with the existing MLE method as well as with the Al-Nasser and Ebrahem (2005) method.

3.3 The Proposed Robust Nonparametric Estimation Method

The following are the steps involved in the proposed method. Firstly, arrange the observed pairs (x_i, y_i)'s, where $i = 1,2,...,n;$ according to the magnitude of x value, by taking into account that all the values of x are distinct. Next, sort these observations into several groups to obtain all the possible paired of slopes. Later on, determine another possible paired of slopes by arranging the observed pairs according to the magnitude of y value. The following are the steps to estimate the slope parameter in LFRM:

Step 1:

Arrange the observations in ascending order, based on x value, i.e., $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$. The associated values of y which may not be in ascending order are taken, i.e., $y_{[1]}, y_{[2]}, ..., y_{[n]}$. The new pairs will be $(x_{(i)}, y_{[i]})$. The i values that are in a bracket, ( ) indicates that they are arranged in ascending order, while i values that are in a square bracket, [ ] indicates that they are not arranged in ascending order.

Step 2:

All the data are divided into $m-$subsamples. These subsamples contains r elements, such that $m \times r = n$. The samples are arranged in the following form:

$$\begin{array}{cccc} (x_{(1)}, y_{[1]}) & (x_{(2)}, y_{[2]}) & \cdots & (x_{(r)}, y_{[r]}) \\ (x_{(r+1)}, y_{[r+1]}) & (x_{(r+2)}, y_{[r+2]}) & \cdots & (x_{(2r)}, y_{[2r]}) \\ \vdots & \vdots & \vdots & \vdots \\ (x_{((m-1)*(r+1))}, y_{[(m-1)*(r+1)]}) & \cdots & \cdots & (x_{(mr)}, y_{[mr]}) \end{array} , \tag{3.1}$$

where m is the maximum divisor of n, such that $m \le r$. As an example, if $n = 50,$ then $m = 5$ and $r = 10$ respectively.

Step 3:

Find all the possible paired slopes.

$$\left\{ b_x(k)_{ij} = \frac{y_{[j]} - y_{[i]}}{x_{(j)} - x_{(i)}}; i = 1,2,\ldots, j-1;\ j = 2,3,\ldots,r \right\}; k = 1,2,\ldots,m \tag{3.2}$$

Step 4:

Repeat Steps 1 to 3 by interchanging y and x to get another possible paired slopes of $b_y(k)_{ij}$.

$$\left\{ b_y(k)_{ij} = \frac{y_{(j)} - y_{(i)}}{x_{[j]} - x_{[i]}}; i = 1,2,\ldots, j-1;\ j = 2,3,\ldots,r \right\}; k = 1,2,\ldots,m \tag{3.3}$$

Step 5:

Find the median of all these slopes.

$$\hat{\beta}_{new} = median \left\{ b_x(k)_{ij}, b_y(k)_{ij} \right\} \tag{3.4}$$

The steps described in Step 1 till Step 3 for estimating the slope parameter is based on the nonparametric estimation method as introduced by Al-Nasser an Ebrahem (2005). In this proposed method, the method is extended by adding two more steps in the procedure to ensure the robustness of this method. In other words, the median of all the slopes when x is arranged in ascending order, and similarly when y is arranged in ascending order is found. The reason why median of all the slopes in Step 5 is used is that the median is found to be more robust than using the mean, when outliers are present in the data (Hampel et al., 2011).

3.4 Simulation Study

A simulation study is performed to compare the proposed method of estimation with the standard MLE method and the nonparametric method by Al-Nasser and Ebrahem (2005). Cases when there are no outlier and also when there are different percentages of outliers are considered. Begin by simulating observations from the LFRM where the parameters are set to $\alpha = 1, \beta = 1$, $\sigma_\delta^2 = 0.1$, and $\lambda = 1$ respectively. The following equations are:

$$Y_i = 1 + X_i, \quad x_i = X_i + \delta_i \text{ and } y_i = Y_i + \varepsilon_i,$$

$$\text{where } X_i = 10\frac{i}{n} \text{ and } \delta_i, \varepsilon_i \sim N(0, 0.1), \text{ with } i = 1, 2, \ldots, n. \tag{3.5}$$

Setting $X_i = 10\frac{i}{n}$, and assuming that X is a mathematical or fixed variable, means that it does not have a specific distribution. If X is a random variable which has a specified distribution then this is termed a structural relationship model between X and Y. Later on the data is contaminated at different levels by replacing the original observation by the contaminated observations. The contaminated observations are generated using the given relationship where $\varepsilon_i \sim N(0, 25)$. The performance of these methods are examined by looking at the mean square error (MSE) of the slope and also the estimated bias (EB) of the parameters in 10,000 trials. The MSE and EB are defined by,

$$MSE = \frac{1}{s}\sum\left(\hat{w}_j - w\right)^2 \text{ and } EB = \left|\hat{w} - w\right|, \tag{3.6}$$

where w is the generic term for the parameters, and s is the sample size.

For each simulation, generate a sample size with $n = 20, 50$ and 100 from the sampling distribution as in (3.5). In order to investigate the robustness of the proposed method, the non-normal error terms are also considered whereby the error terms δ_i and ε_i are generated from three different Beta distributions. The probability density function of the standard Beta distribution can be written as,

$$f(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)} \qquad 0 \le x \le 1;\ p, q > 0\,, \tag{3.7}$$

where p and q are the shape parameters, and $B(p,q)$ can be noted as the Beta function. In this study, the symmetric Beta distribution is considered with parameters (3, 3), right skewed Beta distribution (2, 9) and left skewed Beta distribution (9, 2), respectively using the same above relationship as shown in Figure 3.1.

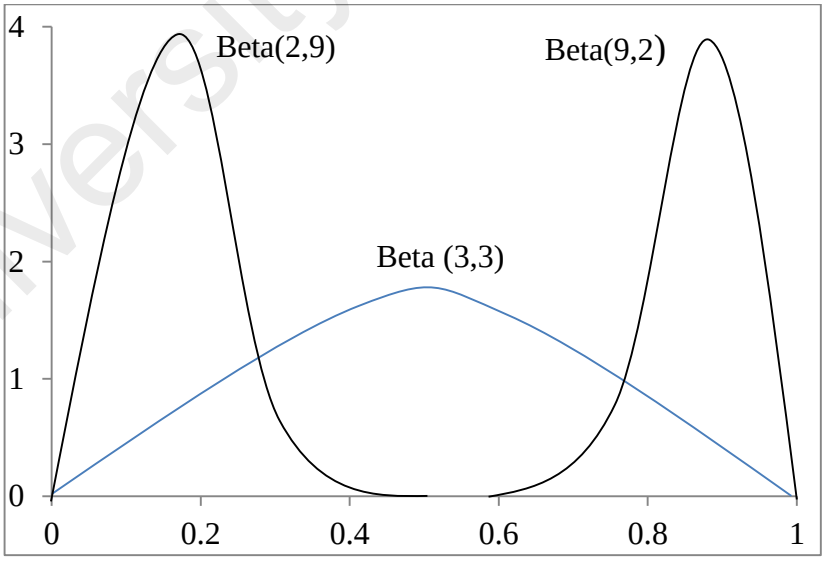

Figure 3.1: Three different non-normal error distribution for δ_i and ε_i

3.5 Results and Discussion

The MSE of the slope parameter are summarised in Table 3.1 to Table 3.4, while the EB of the slope parameter are summarised in Table 3.5 to Table 3.8. Looking at Table 3.1 where the errors δ_i and ε_i are normally distributed, the MSE of the proposed method is somewhat similar to that of the MLE and the nonparametric method by Al-Nasser and Ebrahem (2005) when no outlier exists in the data. However, a great difference can be observed when the data is contaminated. The MSE of the slope estimator using MLE method breaks down easily and becomes huge. Examining closely the proposed method with the nonparametric method by Al-Nasser and Ebrahem (2005) when there is a single outlier, 10%, 20% and 30% outliers, the MSE values of the proposed method has smaller values than the MSE of Al-Nasser and Ebrahem (2005) method. In short, the proposed method outperforms the MLE and the nonparametric method by Al-Nasser and Ebrahem (2005).

Table 3.1: MSE of the slope for normal-case

Contamination	Sample Size / Methods	20	50	100
No outlier	MLE	1.1825E-04	4.7854E-05	2.4419E-05
	Al-Nasser and Ebrahem (2005)	1.5459E-04	5.7457E-05	2.9538E-05
	Proposed	1.5464E-04	5.5672E-05	2.7677E-05
Single outlier	MLE	4.4369E+01	6.4742E-01	8.7298E-02
	Al-Nasser and Ebrahem (2005)	2.2430E-04	7.2124E-05	2.1181E-02
	Proposed	2.2419E-04	6.6974E-05	4.0672E-04
10%	MLE	1.5929E+02	1.6038E+02	1.6043E+02
	Al-Nasser and Ebrahem (2005)	4.8663E-04	4.7335E-04	4.4865E-04
	Proposed	4.8642E-04	4.4584E-04	4.0672E-04
20%	MLE	3.9998E+01	4.0067E+01	4.0083E+01
	Al-Nasser and Ebrahem (2005)	4.3560E-03	3.5335E-03	3.5644E-03
	Proposed	4.3562E-03	3.3497E-03	3.2682E-03
30%	MLE	3.1452E+01	3.1495E+01	3.1498E+01
	Al-Nasser and Ebrahem (2005)	2.6180E+00	3.4945E-02	3.6090E-02
	Proposed	2.6179E+00	3.1784E-02	3.1157E-02

Meanwhile, for Table 3.2 where the errors δ_i and ε_i are skewed to the right with Beta distribution (2, 9), the MSE of the proposed method also show similar results to that of the MLE and the nonparametric method by Al-Nasser and Ebrahem (2005) in the case when no outlier is present. When the data gets contaminated, the MSE of the slope estimator using MLE method breaks down

easily. However, the MSE values for the proposed method and nonparametric method by Al-Nasser and Ebrahem (2005) are not affected by the outliers even when the percentage of outliers increase. Comparing the proposed method with Al-Nasser and Ebrahem (2005) method, at each level of contamination, the proposed method shows consistently smaller values of MSE than that of Al-Nasser and Ebrahem (2005) method.

Table 3.2: MSE of the slope for right skewed case, Beta (2, 9)

Contamination	**Sample Size** / **Methods**	**20**	**50**	**100**
No outlier	MLE	1.5133E-04	6.0594E-05	3.0176E-05
	Al-Nasser and Ebrahem (2005)	1.9075E-04	6.8761E-05	3.4546E-05
	Proposed	1.9064E-04	6.6482E-05	3.2149E-05
Single outlier	MLE	4.4529E+01	6.4764E-01	8.7356E-02
	Al-Nasser and Ebrahem (2005)	2.7768E-04	8.6845E-05	4.0448E-05
	Proposed	2.7693E-04	7.9988E-05	3.4972E-05
10%	MLE	1.5966E+02	1.6054E+02	1.6055E+02
	Al-Nasser and Ebrahem (2005)	6.0344E-04	5.7011E-04	5.3175E-04
	Proposed	6.0091E-04	5.2728E-04	4.7306E-04
20%	MLE	4.0002E+01	4.0076E+01	4.0081E+01
	Al-Nasser and Ebrahem (2005)	5.3034E-03	4.3112E-03	4.2902E-03
	Proposed	5.2779E-03	4.0164E-03	3.8901E-03
30%	MLE	3.1461E+01	3.1494E+01	3.1501E+01
	Al-Nasser and Ebrahem (2005)	2.6401E+00	4.3176E-02	4.3854E-02
	Proposed	2.6394E+00	3.8207E-02	3.7091E-02

Next, from Table 3.3, where the errors δ_i and ε_i are skewed to the left with Beta distribution (9, 2) the MSE obtained are consistent for the MLE method, the nonparametric method by Al-Nasser and Ebrahem (2005) and the proposed method when no outlier is present in the data. However, when there is a single outlier, 10%, 20% or 30% outlier in the data set, the MSE for MLE method becomes very huge. The MSE for the proposed method and the nonparametric method by Al-Nasser and Ebrahem (2005), on the other hand are consistently small and has no effect in the presence of the outliers. A closer look of the MSE values for both these two methods show that the proposed method gave smaller values than the method proposed by Al-Nasser and Ebrahem (2005).

Table 3.3: MSE of the Slope for left skewed case, Beta (9, 2)

Contamination	**Sample Size** / **Methods**	**20**	**50**	**100**
No outlier	MLE	1.5120E-04	6.0583E-05	3.0172E-05
	Al-Nasser and Ebrahem (2005)	1.9106E-04	6.8610E-05	3.4436E-05
	Proposed	1.9089E-04	6.5936E-05	3.1775E-05
Single outlier	MLE	4.4658E+01	6.4804E-01	8.7271E-02
	Al-Nasser and Ebrahem (2005)	2.7683E-04	8.6440E-05	4.0811E-05
	Proposed	2.7580E-04	7.9212E-05	3.5052E-05
10%	MLE	1.5982E+02	1.6046E+02	1.6037E+02
	Al-Nasser and Ebrahem (2005)	5.9557E-04	5.6534E-04	5.3778E-04
	Proposed	5.9307E-04	5.2404E-04	4.7994E-04
20%	MLE	3.9997E+01	4.0071E+01	4.0081E+01
	Al-Nasser and Ebrahem (2005)	5.3072E-03	4.2962E-03	4.3022E-03
	Proposed	5.2842E-03	4.0062E-03	3.9089E-03

30%	MLE	3.1455E+01	3.1498E+01	3.1499E+01
	Al-Nasser and Ebrahem (2005)	2.6396E+00	4.2998E-02	4.3809E-02
	Proposed	2.6388E+00	3.8164E-02	3.7132E-02

For Table 3.4, with errors δ_i and ε_i that are non-normal symmetric case with Beta distribution (3, 3) all three methods show somewhat similar MSE values when no outlier exist in the data. However, when there are outliers in the data, the MSE of the slope for MLE method breaks down quickly and becomes huge. The MSE of the slope for the proposed method and the nonparametric method by Al-Nasser and Ebrahem (2005), on the other hand, remain small and are not affected by the presence of outliers. Comparing the MSE of the nonparametric method by Al-Nasser and Ebrahem (2005) with the proposed method, it can be observed that the proposed method gives a more satisfactory result in the presence of outliers by having smaller values of MSE.

Table 3.4: MSE of the Slope for non-normal symmetric case, Beta (3, 3)

Contamination	Sample Size / Methods	20	50	100
No outlier	MLE	4.1847E-04	1.7195E-04	8.6134E-05
	Al-Nasser and Ebrahem (2005)	5.7502E-04	2.2126E-04	1.1804E-04
	Proposed	5.6131E-04	2.0281E-04	1.0113E-04
Single outlier	MLE	4.5718E+01	6.4956E-01	8.7330E-02
	Al-Nasser and Ebrahem (2005)	8.6448E-04	2.9515E-04	1.4889E-04
	Proposed	8.1974E-04	2.4402E-04	1.1237E-04
10%	MLE	1.6275E+02	1.6150E+02	1.6107E+02
	Al-Nasser and Ebrahem (2005)	1.8795E-03	1.8761E-03	1.8190E-03
	Proposed	1.7693E-03	1.5884E-03	1.4830E-03
20%	MLE	4.0043E+01	4.0100E+01	4.0091E+01
	Al-Nasser and Ebrahem (2005)	1.5717E-02	1.3294E-02	1.3519E-02
	Proposed	1.4875E-02	1.1492E-02	1.1323E-02
30%	MLE	3.1459E+01	3.1504E+01	3.1501E+01
	Al-Nasser and Ebrahem (2005)	2.7694E+00	1.2712E-01	1.2854E-01
	Proposed	2.7570E+00	9.9070E-02	9.6533E-02

From Table 3.1 to Table 3.4, it can be seen that the MLE method breaks down easily when outliers are present in the data. In contrast, the nonparametric method proposed by Al-Nasser and Ebrahem (2005) and the proposed method have a more satisfactory result in the presence of outliers. However, the MSE values shows that the proposed method gives a consistently smaller values than the nonparametric method by Al-Nasser and Ebrahem (2005).

In addition to the MSE of the slope for each method, the EB of the slope parameter after 10,000 simulation process is measured. The results are shown in Table 3.5 to Table 3.8 respectively.

Table 3.5: EB of the slope: Normal-Case

Contamination	**Sample size** / **Methods**	**20**	**50**	**100**
No outlier	MLE	1.9350E-04	9.7890E-06	2.3213E-05
	Al-Nasser and Ebrahem (2005)	2.8836E-04	4.4696E-04	7.1954E-04
	Proposed	2.9352E-04	2.0146E-05	4.3393E-05
Single outlier	MLE	6.6469E+00	8.0430E-01	2.9533E-01
	Al-Nasser and Ebrahem (2005)	6.8065E-03	3.3655E-03	2.0164E-02
	Proposed	6.8007E-03	2.8622E-03	1.9165E-02
10%	MLE	1.2604E+01	1.2657E+01	1.2662E+01
	Al-Nasser and Ebrahem (2005)	1.6510E-02	1.9768E-02	2.0164E-02
	Proposed	1.6504E-02	1.9161E-02	1.9165E-02
20%	MLE	6.3234E+00	6.3293E+00	6.3306E+00
	Al-Nasser and Ebrahem (2005)	6.1905E-02	5.7962E-02	5.8935E-02
	Proposed	6.1905E-02	5.6511E-02	5.6492E-02
30%	MLE	5.6076E+00	5.6116E+00	5.6120E+00
	Al-Nasser and Ebrahem (2005)	1.6177E+00	1.8411E-01	1.8853E-01
	Proposed	1.6177E+00	1.7623E-01	1.7557E-01

Table 3.6: EB of the slope: Right skewed case, Beta (2, 9)

Contamination	**Sample size** / **Methods**	**20**	**50**	**100**
No outlier	MLE	5.0579E-05	3.5651E-05	1.2450E-04
	Al-Nasser and Ebrahem (2005)	3.0781E-05	6.2823E-04	8.4873E-04
	Proposed	3.9226E-07	1.3998E-05	1.1493E-04
Single outlier	MLE	6.6558E+00	8.0437E-01	2.9540E-01
	Al-Nasser and Ebrahem (2005)	7.7383E-03	3.8063E-03	2.4013E-03
	Proposed	7.7029E-03	3.1258E-03	1.4051E-03
10%	MLE	1.2614E+01	1.2662E+01	1.2666E+01
	Al-Nasser and Ebrahem (2005)	1.8366E-02	2.1643E-02	2.1941E-02
	Proposed	1.8318E-02	2.0764E-02	2.0655E-02
20%	MLE	6.3236E+00	6.3299E+00	6.3305E+00
	Al-Nasser and Ebrahem (2005)	6.8161E-02	6.3939E-02	6.4639E-02
	Proposed	6.8031E-02	6.1831E-02	6.1615E-02
30%	MLE	5.6084E+00	5.6115E+00	5.6122E+00
	Al-Nasser and Ebrahem (2005)	1.6245E+00	2.0457E-01	2.0769E-01
	Proposed	1.6242E+00	1.9321E-01	1.9149E-01

Table 3.7: EB of the slope: Left skewed case, Beta (9, 2)

Contamination	**Sample size** / **Methods**	**20**	**50**	**100**
No outlier	MLE	1.0542E-04	2.6538E-05	8.8454E-05
	Al-Nasser and Ebrahem (2005)	1.6405E-04	5.8414E-04	1.0377E-03
	Proposed	1.9086E-04	4.3450E-05	8.8449E-05
Single outlier	MLE	6.6646E+00	8.0462E-01	2.9526E-01
	Al-Nasser and Ebrahem (2005)	7.4805E-03	3.7438E-03	2.5641E-03
	Proposed	7.4449E-03	3.0662E-03	1.5910E-03
10%	MLE	1.2622E+01	1.2659E+01	1.2659E+01
	Al-Nasser and Ebrahem (2005)	1.8092E-02	2.1578E-02	2.2095E-02
	Proposed	1.8052E-02	2.0722E-02	2.0835E-02
20%	MLE	6.3232E+00	6.3296E+00	6.3305E+00
	Al-Nasser and Ebrahem (2005)	6.8118E-02	6.3858E-02	6.4751E-02
	Proposed	6.7989E-02	6.1780E-02	6.1778E-02
30%	MLE	5.6078E+00	5.6119E+00	5.6120E+00
	Al-Nasser and Ebrahem (2005)	1.6243E+00	2.0408E-01	2.0763E-01
	Proposed	1.6240E+00	1.9305E-01	1.9160E-01

Table 3.8: EB of the slope: Non-Normal Symmetric case, Beta (3, 3)

Contamination	**Sample size** / **Methods**	**20**	**50**	**100**
No outlier	MLE	1.1409E-04	3.9695E-05	4.2878E-05
	Al-Nasser and Ebrahem (2005)	1.0258E-03	2.6493E-03	3.4802E-03
	Proposed	1.6671E-04	2.2886E-05	1.8443E-04
Single outlier	MLE	6.7108E+00	8.0489E-01	2.9508E-01
	Al-Nasser and Ebrahem (2005)	1.4711E-02	8.3751E-03	6.2885E-03
	Proposed	1.3715E-02	5.5138E-03	2.9021E-03
10%	MLE	1.2697E+01	1.2684E+01	1.2679E+01
	Al-Nasser and Ebrahem (2005)	3.3244E-02	3.9754E-02	4.0872E-02
	Proposed	3.1964E-02	3.6266E-02	3.6692E-02
20%	MLE	6.3254E+00	6.3312E+00	6.3309E+00
	Al-Nasser and Ebrahem (2005)	1.1801E-01	1.1276E-01	1.1498E-01
	Proposed	1.1514E-01	1.0498E-01	1.0532E-01
30%	MLE	5.6076E+00	5.6122E+00	5.6121E+00
	Al-Nasser and Ebrahem (2005)	1.6633E+00	3.5140E-01	3.5582E-01
	Proposed	1.6596E+00	3.1183E-01	3.0923E-01

These simulation study includes cases when the errors δ_i and ε_i are normally distributed, skewed to the right with Beta distribution (2, 9), skewed to the left with Beta distribution (9, 2), and non-normal symmetric case with Beta distribution (3, 3) respectively. Results from all the four cases mentioned show that when there are no outliers, all the methods of estimation show somewhat similar results for EB values.

However, as the percentage of contamination increases, the EB values for the MLE method increase dramatically. Comparing the proposed method with Al-Nasser and Ebrahem (2005) method, note that the proposed method always gave smaller EB values in situation when there are no outliers, as well as in situation when there is a single, 10%, 20%, and 30% outliers.

To summarise, the extension on the nonparametric method by Al-Nasser and Ebrahem (2005) in estimating the slope parameter is considered a robust method as it gives a more satisfactory result in terms of small MSE values and small EB values. The extension made which includes arranging the data according to the magnitude of y observations as well as finding the median of all the slopes, result in a more robust estimation even with the existence of high percentage of outliers.

3.6 Practical Example

In this section, the proposed nonparametric technique is applied to real life data and the three estimation methods namely the MLE method, the nonparametric method by Al-Nasser and Ebrahem (2005), and the proposed method are compared. By considering a real data set from a study conducted by Goran et al. (1996), the data set comprises of 96 observations that are free from any outliers. The study was to examine the accuracy of some widely used body-composition techniques for children between the ages of 4 and 10 years by two different techniques, namely skinfold thickness (ST) and bioelectrical resistance (BR). Measurement error are assumed to occur in either variable of this experiment to make the relationship as given in (2.1).

In order to apply the proposed method to estimate the slope parameter, these data sets are divided into 8 groups, with each group having 12 elements. Next, to examine the

effect on the slope with the presence of outlying observation, Goran et al. (1996) data is modified by inserting several outliers to create different outlier situation. The outliers are inserted by following Kim (2000) and Imon and Hadi (2008) where a certain percentage of the observations are removed and replaced with the outliers' observation. The contaminated observation were generated based on the given relationship where $\varepsilon_i \sim N(0, 25)$. In this study, these cases are being considered; when there is a single outlier, 10%, 20%, and 30% outliers respectively. The estimated slopes by using three different methods are shown in Table 3.9.

Table 3.9: The Slope Estimates using Three Different Methods from Goran et al. (1996)

Contamination	Methods	Slopes	Standard deviation
No outlier	MLE	1.0988	0.0576
	Al-Nasser and Ebrahem (2005)	1.0016	0.0544
	Proposed	1.0268	0.0548
Single outlier	MLE	3.7150	1.8907
	Al-Nasser and Ebrahem (2005)	1.0093	1.0737
	Proposed	1.0274	1.0691
10%	MLE	21.3319	37.5925
	Al-Nasser and Ebrahem (2005)	1.0274	10.8514
	Proposed	1.0579	10.7201
20%	MLE	27.9859	57.5872
	Al-Nasser and Ebrahem (2005)	1.0961	15.1723
	Proposed	1.1056	15.1074
30%	MLE	49.5205	175.3813
	Al-Nasser and Ebrahem (2005)	1.0806	28.3873
	Proposed	1.1011	28.1275

From Table 3.9 when there are no outliers in the data set, all the three methods show somewhat similar results in terms of the slope parameter. However, as the percentage of outlier increases, the slope for MLE method changes significantly compared to when there is no outlier. To conclude, the MLE estimation method breaks down easily with the increase of percentage of outliers, as compared to the nonparametric

method by Al-Nasser and Ebrahem (2005) and the proposed method. Looking closely to the slopes parameter of the nonparametric method by Al-Nasser and Ebrahem (2005) and the proposed method, both these estimation methods are not much affected by the presence of outliers. However, comparing the standard deviation of the nonparametric method by Al-Nasser and Ebrahem (2005) and the proposed method when outliers exist, this proposed method results in a lower value of standard deviation, which indicates that the data are clustered close to the mean, thus being more reliable.

3.7 Summary

In this chapter, a robust nonparametric method to estimate the slope parameter of the LFRM is proposed. From the simulation study, all the three methods give somewhat similar results when no outlier exists in the data. However, as the percentage of outlier increases, the MLE method shows to break down quickly as compared to the nonparametric method by Al-Nasser and Ebrahem (2005) and the proposed method. Comparing the MSE and the EB of the nonparametric method (Al-Nasser and Ebrahem, 2005) and the proposed method, it can be summarized that the proposed method gives a more satisfactory result in the presence of outliers.

The applicability of these methods are illustrated in a real data set by Goran et al. (1996) and to conclude, when no outliers exist in the data set, all the three methods show somewhat similar results in terms of slope parameter; however, as the percentage of outlier increases, the MLE method shows to break down easily. Comparing Al-Nasser and Ebrahem (2005) and the proposed method, it can be concluded that the slope parameters are not much affected by the presence of outliers, and based on the standard deviation of both methods, the proposed method shows to be more accurate by having smaller values of standard deviation as compared to the nonparametric method by Al-Nasser and Ebrahem (2005) at each percentage of data missing. To summarise, the

proposed method to estimate the slope parameter of the LFRM which is an extension from Al-Nasser and Ebrahem (2005) nonparametric method, is considered as the best method as it shows to perform well in the presence of small percentage of outliers as well as high percentage of outliers.

CHAPTER 4: SINGLE OUTLIER DETECTION USING COVRATIO STATISTIC

4.1 Introduction

This chapter derives a test statistic based on *COVRATIO* to detect an outlier in the LFRM. Section 4.2 describes the *COVRATIO* statistic for LFRM. Next, the procedure for determining the cut-off points for detecting an outlier is given in Section 4.3. Section 4.4 covers the measure of power performance of the *COVRATIO* statistic in determining an outlier. Section 4.5 illustrates the applicability of the *COVRATIO* statistic using a practical example, followed by a real data illustration in Section 4.6. Finally, a summary of the chapter is given in Chapter 4.7.

4.2 *COVRATIO* Statistic for Linear Functional Relationship Model

The analysis of errors-in-variable model (EIVM) may be subjected to the occurrence of outliers (Hadi et al., 2009). Several researchers have studied the outliers problem for different models; for example Belsley et al. (1980) and Barnet and Lewis (1994) have thoroughly discussed the problem of outliers in a linear regression. Abuzaid et al. (2011) on the other hand, discussed the identification of outliers in a circular regression by using a new definition of circular residuals via different graphical and numerical methods. Recently, Ibrahim et al. (2013) proposed the *COVRATIO* procedure to detect outliers in a circular regression model.

The *COVRATIO* procedure dates back to Belsley et al. (1980) in which they proposed a numerical statistic to identify outliers in a linear regression models. This numerical 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. If the ratio is close to 1, then there is no significant difference between them. In other words, the i^{th} observation

is consistent with the other observations. Alternatively, if the value of $\left|COVRATIO_{(-i)} - 1\right|$ is close to or larger than $\left(\frac{3p}{n}\right)$, then it indicates that the i^{th} observation is a candidate of an outlier, where p is the number of parameters, and n is the sample size. In this chapter, this idea is extended to the LFRM in which the cut-off values and formulas will be derived.

As a starting point, the determinantal ratio is applied to the LFRM to obtain the cut-off point. The determinant of the covariance matrix is given by

$$|COV| = \hat{Var}(\hat{\alpha})\hat{Var}(\hat{\beta}) - \hat{Cov}(\hat{\alpha}, \hat{\beta})^2 \tag{4.1}$$

where

$$\hat{Var}(\hat{\alpha}) = \frac{(\lambda + \hat{\beta}^2)\hat{\sigma}_\delta^2 \hat{\beta}}{S_{xy}} \left\{ \bar{x}^2 (1 + \hat{T}) + \frac{S_{xy}}{n\hat{\beta}} \right\},$$

$$\hat{Var}(\hat{\beta}) = \frac{(\lambda + \hat{\beta}^2)\hat{\sigma}_\delta^2 \hat{\beta}}{S_{xy}} \left\{ 1 + \hat{T} \right\},$$

$$\hat{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{(\lambda + \hat{\beta}^2)\hat{\sigma}_\delta^2 \hat{\beta}\,\bar{x}}{S_{xy}} \left\{ 1 + \hat{T} \right\},$$

and $$\hat{T} = \frac{n\lambda\hat{\beta}\hat{\sigma}_\delta^2}{(\lambda + \hat{\beta}^2)S_{xy}}.$$

The $COVRATIO$ statistic for the i^{th} observation is given by

$$COVRATIO_{(-i)} = \frac{\left|COV\right|}{\left|COV_{(-i)}\right|}, \tag{4.2}$$

whereby $\left|COV\right|$ is the

covariance matrix for the full data set and $\left|COV_{(-i)}\right|$ is the corresponding covariance matrix after deleting the i^{th} observation. Any observation with $\left|COVRATIO_{(-i)} - 1\right|$ that exceeds the cut-off points will be identified as an outlier. The cut-off points are obtained through simulation studies.

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

The cut-off points of the $COVRATIO$ statistic to identify the outliers in the LFRM are obtained by applying Monte Carlo simulation method. Fifteen different sample sizes of $n = 30, 40, 50, 60, ..., 150, 250$ and 500, and five values of σ_ε = 0.2, 0.4, 0.6, 0.8, and 1.0 respectively are used. For each sample of size n and σ_ε, a set of normal random errors are generated from the normal distribution with mean 0 and σ_ε respectively. Assume both $\delta_i = \varepsilon_i$, and proceed with the following steps:

Step 1. Generate $X_i = 10\left(\frac{i}{n}\right)$ of size n, with $i = 1,2,3,...,n,$ where n is the sample size. Without loss of generality, the slope and intercept parameters of LFRM are fixed at $\beta = 1$ and $\alpha = 0$ respectively.

Step 2. Generate two random error terms δ_i and ε_i of size *n* from $N(0, \sigma_\delta^2)$ and $N(0, \sigma_\varepsilon^2)$ respectively, as in (2.3).

Step 3. Calculate the observed values of x_i and y_i using (2.2).

Step 4. Fit the generated data by using the parametric LFRM from (2.2).

Step 5. Calculate $|COV|$ by using (4.1).

Step 6. Exclude the i^{th} row from the generated sample, where $i = 1, 2, 3, \ldots, n$. Then for all i, repeat steps 4 till step 6 to obtain $|COV_{(-i)}|$.

Step 7. Calculate $COVRATIO_{(-i)}$ by using (4.2) and find the value of $|COVRATIO_{(-i)} - 1|$ for all i.

Step 8. State the maximum value of $|COVRATIO_{(-i)} - 1|$.

These steps are repeated for 10,000 times for each combination of sample size n and σ_ε. After that, the 1%, 5%, and 10% upper percentiles of the maximum values of $|COVRATIO_{(-i)} - 1|$ are calculated. These upper percentiles are used as the cut-off points in identifying the outliers for the LFRM. The R programming code for this simulation process is given in Appendix B.

The plots in Figure 4.1 to Figure 4.6 respectively gives the 1%, 5%, and 10% upper percentile values of $|COVRATIO_{(-i)} - 1|$ against σ_ε, for selected n namely $n = 50, 70, 100, 150, 250,$ and 500. For other sample sizes, $n = 60, 80, 90, 110, 120, 130$ and 140, the plots of the 1%, 5%, and 10% upper percentile values of $|COVRATIO_{(-i)} - 1|$ against σ_ε can be found in Appendix C. From these figures, it is noted that the 1%, 5% and 10% upper percentile values of $|COVRATIO_{(-i)} - 1|$ are independent of σ_ε for all n.

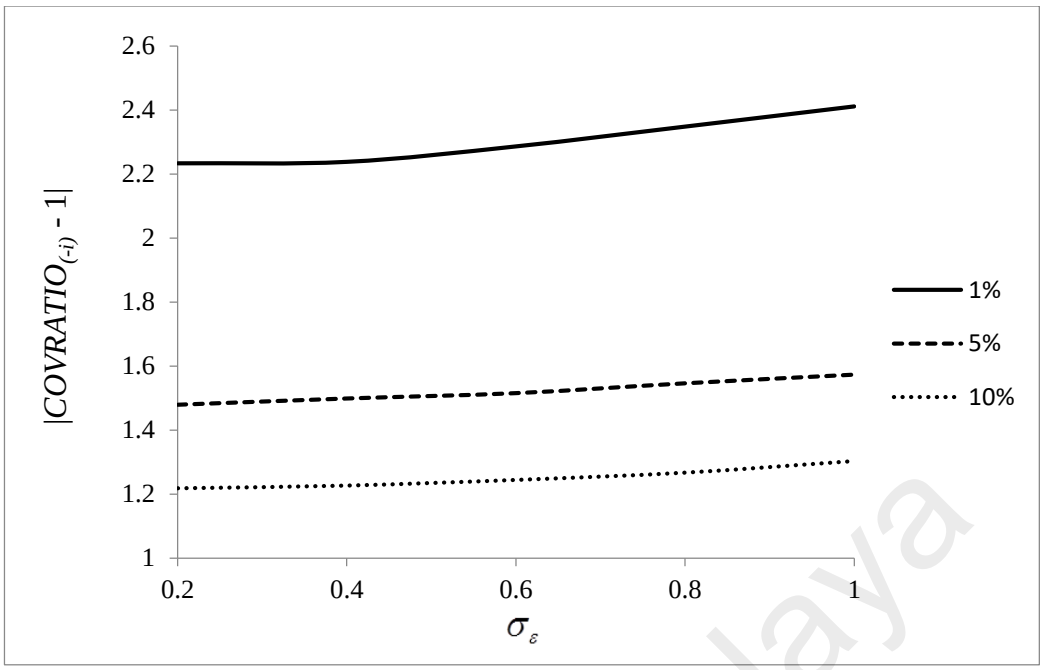

Figure 4.1: The upper percentile points of $\left|COVRATIO_{(-i)}-1\right|$ for $n=50$

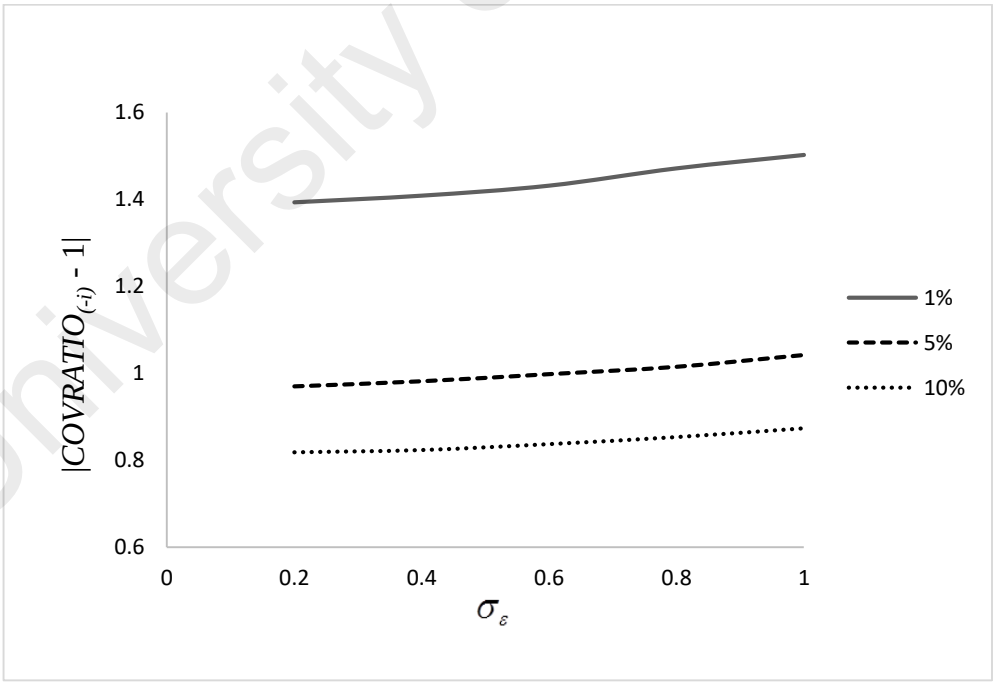

Figure 4.2: The upper percentile points of $\left|COVRATIO_{(-i)}-1\right|$ for $n=70$

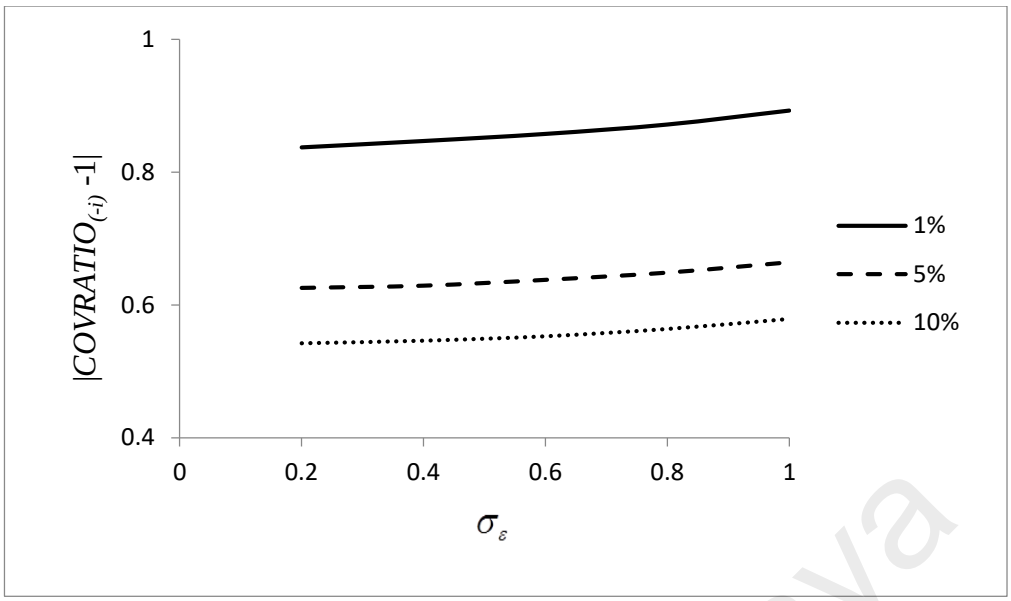

Figure 4.3: The upper percentile points of $\left|COVRATIO_{(-i)}-1\right|$ for $n=100$

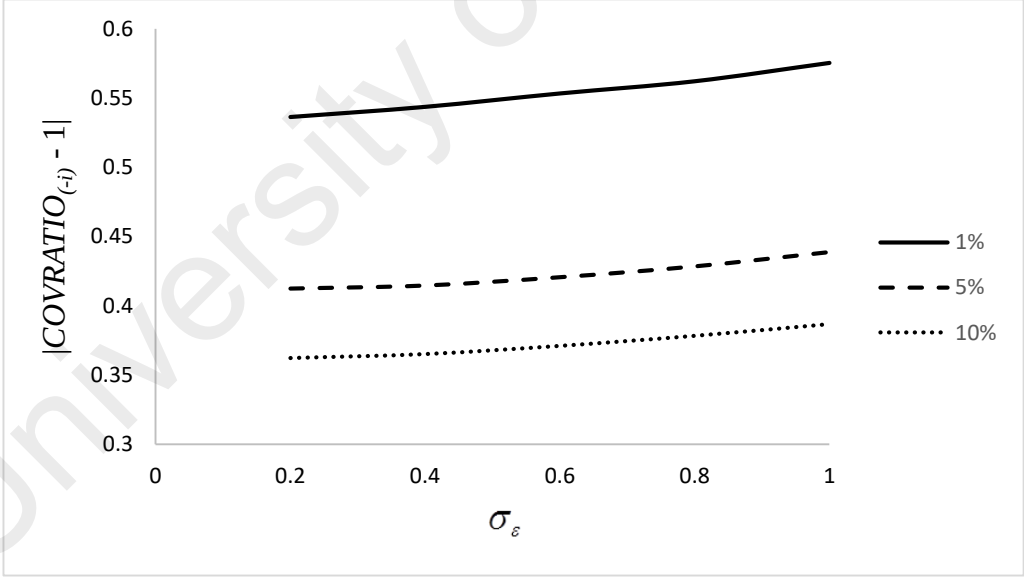

Figure 4.4: The upper percentile points of $\left|COVRATIO_{(-i)}-1\right|$ for $n=150$

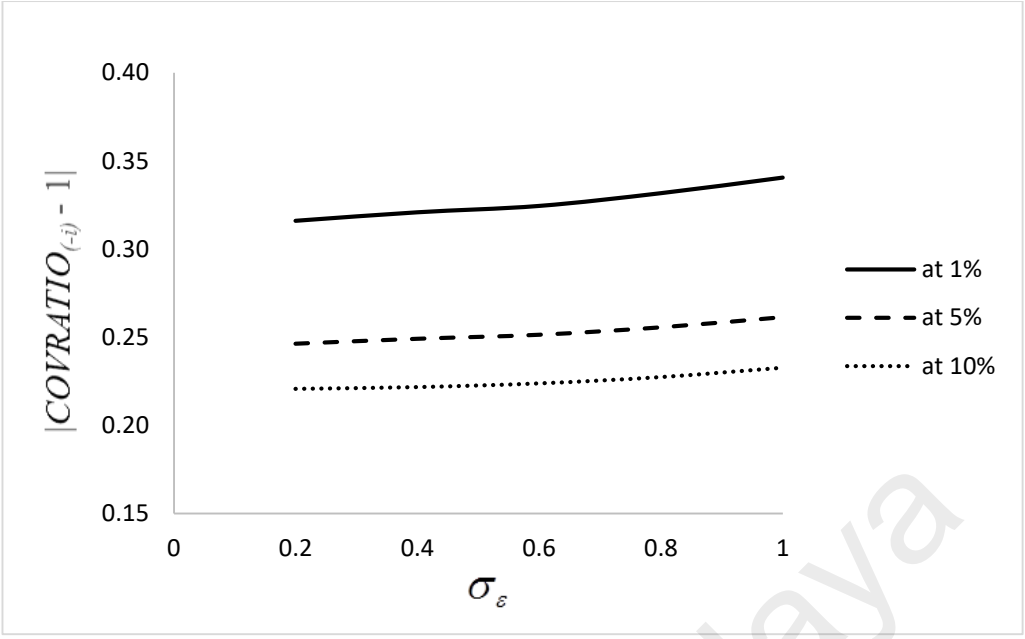

Figure 4.5: The upper percentile points of $\left|COVRATIO_{(-i)}-1\right|$ for $n=250$

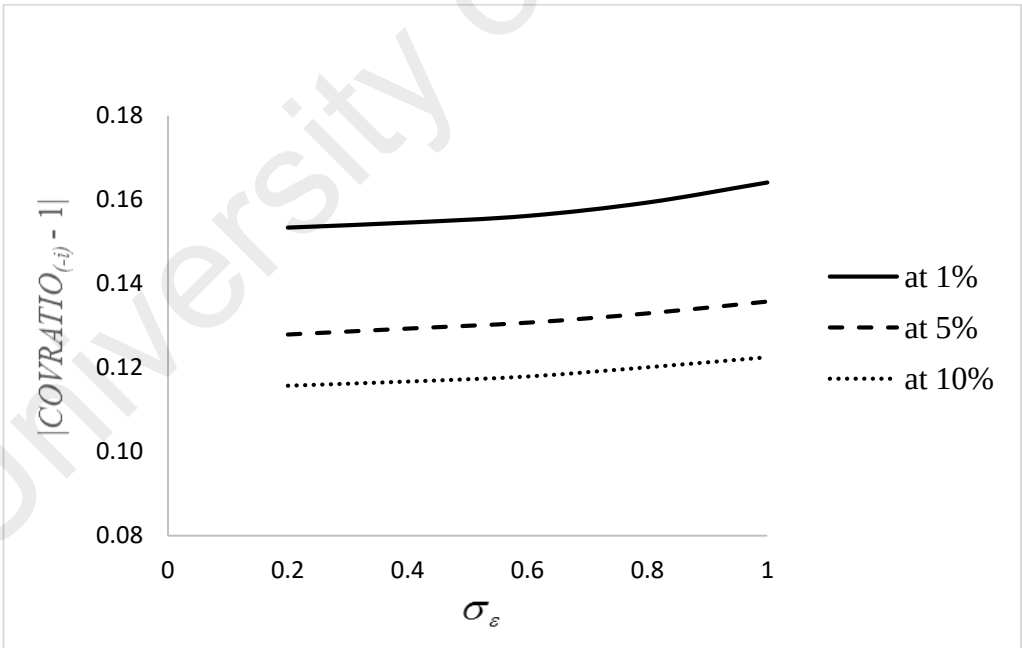

Figure 4.6: The upper percentile points of $\left|COVRATIO_{(-i)}-1\right|$ for $n=500$

Alternatively, the results can be tabulated at different levels of significant. Table 4.1 gives the cut-off points for various sample sizes of n, at 1% significant level while Table 4.2 gives the cut-off points for various sample sizes at 5% significant level. Table 4.3 shows the cut-off points for various sample sizes of n, at 10% significant level. From all the three tables, the cut-off points show a decreasing function of sample size n.

Table 4.1: The 1% upper percentile points of $\left|COVRATIO_{(-i)} - 1\right|$,

at σ_ε = 0.2, 0.4, 0.6, 0.8 & 1.0.

n	**1% at** σ_ε **= 0.2**	**1% at** σ_ε **= 0.4**	**1% at** σ_ε **= 0.6**	**1% at** σ_ε **= 0.8**	**1% at** σ_ε **= 1.0**
30	5.6119	5.6667	5.8076	5.9329	6.2449
40	3.2652	3.2951	3.3952	3.4786	3.6240
50	2.2340	2.2385	2.2868	2.3487	2.4118
60	1.6404	1.6456	1.6940	1.7417	1.8106
70	1.3935	1.4088	1.4318	1.4715	1.5024
80	1.1327	1.1428	1.1666	1.1891	1.2179
90	0.9556	0.9615	0.9812	1.0137	1.0353
100	0.8373	0.8469	0.8576	0.8718	0.8928
110	0.7505	0.7537	0.7556	0.7732	0.7851
120	0.7013	0.7076	0.7157	0.7315	0.7495
130	0.6434	0.6532	0.6569	0.6738	0.6926
140	0.5813	0.5877	0.5960	0.6044	0.6147
150	0.5364	0.5437	0.5533	0.5621	0.5753
250	0.3161	0.3209	0.3245	0.3317	0.3405
500	0.1533	0.1545	0.1561	0.1593	0.1641

Table 4.2: The 5% upper percentile points of $\left|COVRATIO_{(-i)} - 1\right|$,

at σ_ε = 0.2, 0.4, 0.6, 0.8 & 1.0.

n	**5% at** σ_ε **= 0.2**	**5% at** σ_ε **= 0.4**	**5% at** σ_ε **= 0.6**	**5% at** σ_ε **= 0.8**	**5% at** σ_ε **= 1.0**
30	3.3150	3.3731	3.4121	3.4876	3.5920
40	2.0774	2.1014	2.1265	2.1702	2.2423
50	1.4795	1.4991	1.5160	1.5468	1.5740
60	1.1647	1.1750	1.1926	1.2136	1.2386
70	0.9699	0.9817	0.9980	1.0150	1.0421
80	0.8196	0.8277	0.8392	0.8529	0.8744
90	0.7165	0.7202	0.7303	0.7370	0.7533
100	0.6260	0.6294	0.6381	0.6490	0.6646
110	0.5659	0.5711	0.5774	0.5914	0.6053
120	0.5240	0.5286	0.5363	0.5454	0.5577
130	0.4788	0.4848	0.4902	0.4985	0.5110
140	0.4450	0.4498	0.4576	0.4637	0.4739
150	0.4124	0.4146	0.4206	0.4284	0.4388
250	0.2463	0.2491	0.2514	0.2556	0.2614
500	0.1279	0.1293	0.1307	0.1329	0.1357

Table 4.3: The 10% upper percentile points of $\left|COVRATIO_{(-i)} - 1\right|$,

at σ_ε = 0.2, 0.4, 0.6, 0.8 & 1.0.

n	**10% at σ_ε = 0.2**	**10% at σ_ε = 0.4**	**10% at σ_ε = 0.6**	**10% at σ_ε = 0.8**	**10% at σ_ε = 1.0**
30	2.5623	2.5694	2.6108	2.6740	2.7568
40	1.6616	1.6767	1.7001	1.7407	1.7828
50	1.2189	1.2273	1.2449	1.2677	1.3039
60	0.9709	0.9799	0.9941	1.0118	1.0404
70	0.8183	0.8235	0.8370	0.8532	0.8732
80	0.6957	0.7028	0.7136	0.7258	0.7428
90	0.6140	0.6186	0.6232	0.6332	0.6484
100	0.5427	0.5465	0.5531	0.5640	0.5789
110	0.4941	0.5002	0.5058	0.5155	0.5263
120	0.4565	0.4605	0.4665	0.4756	0.4853
130	0.4168	0.4209	0.4250	0.4329	0.4413
140	0.3902	0.3928	0.3975	0.4033	0.4124
150	0.3624	0.3653	0.3712	0.3784	0.3869
250	0.2207	0.2217	0.2238	0.2274	0.2327
500	0.1157	0.1167	0.1179	0.1201	0.1225

Recall for the linear regression model, Belsley et al. (1980) defined $\left(\frac{3p}{n}\right)$ as the cut-off formula for $\left|COVRATIO_{(-i)} - 1\right|$ statistic at 5% significant level, where p is the number of parameters, and n is the sample size. This idea is extended to find the cut-off points for LFRM. Averaging the values of $\left|COVRATIO_{(-i)} - 1\right|$ for each σ_ε at 1%, 5%, and 10%, the curves are plotted as shown in Figure 4.7 to Figure 4.9. Fitting the curve with the power series equation by finding the least square, the equation of the series trend line is obtained as $y = 135.63n^{-1.145}$ for 5% significant level. Similar formulations of

the trend lines are obtained for the 1% and 10% significant level and they are presented in Table 4.4.

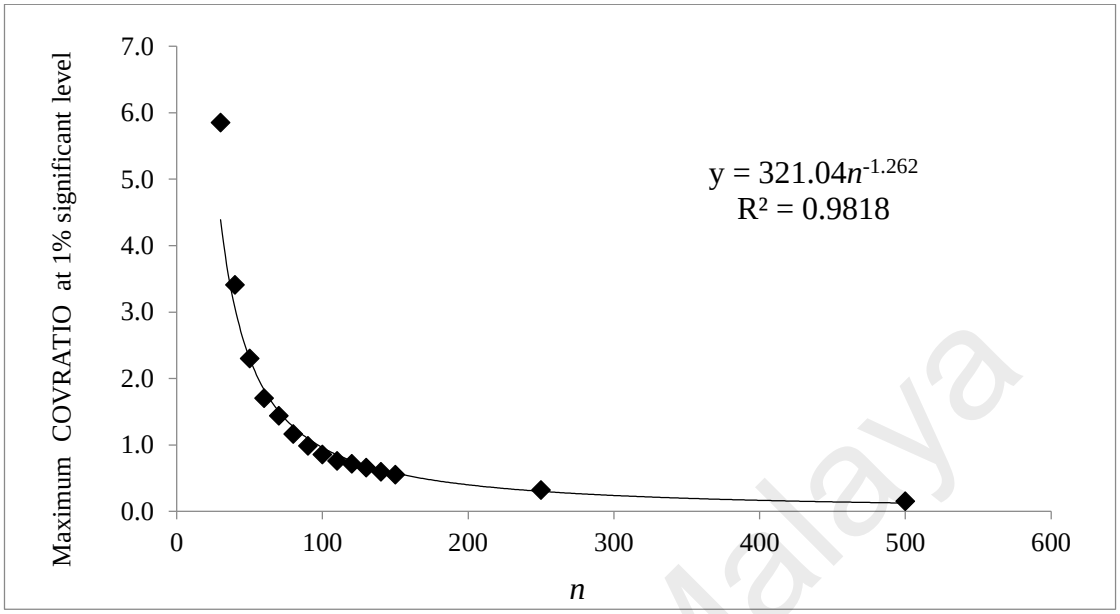

Figure 4.7: Graph of the Power Series in Finding the General Formula for the Cut-Off Point at 1% Significant Level.

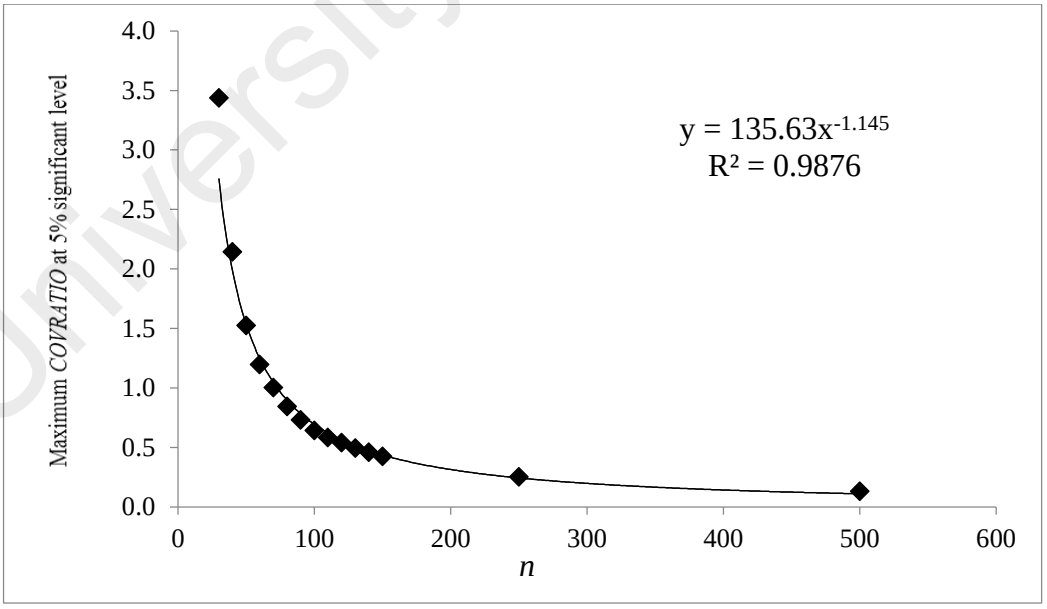

Figure 4.8: Graph of the Power Series in Finding the General Formula for the Cut-Off Point at 5% Significant Level

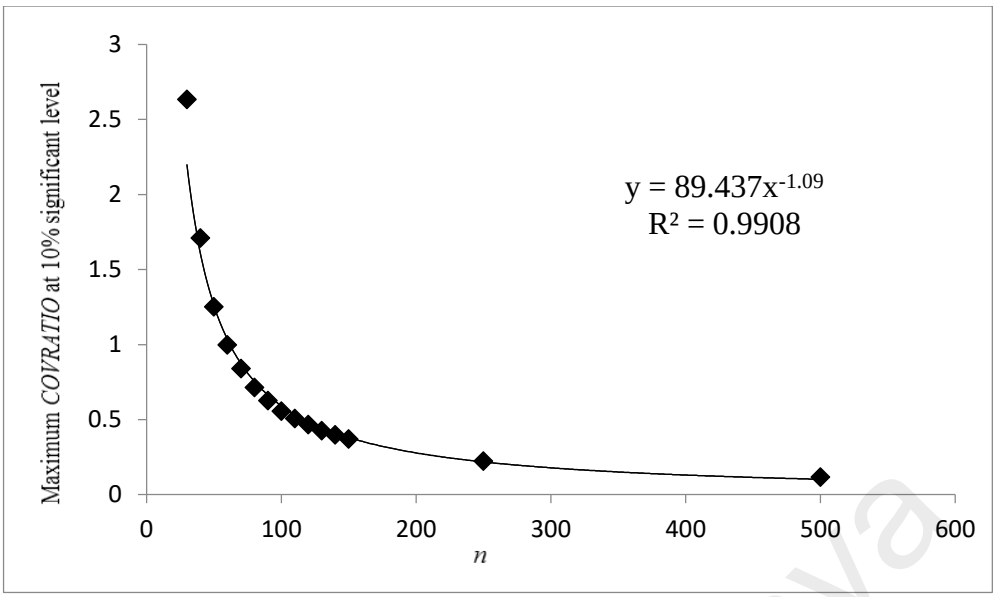

Figure 4.9: Graph of the Power Series in Finding the General Formula for the Cut-Off Point at 10% Significant Level.

Table 4.4: 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**
1%	$y = 321.04n^{-1.262}$
5%	$y = 135.63n^{-1.145}$
10%	$y = 89.44n^{-1.090}$

4.4 Power of Performance for *COVRATIO* Statistic

In detecting an outlier in the LFRM, the performance of $\left|COVRATIO_{(-i)}-1\right|$ is examined by applying Monte Carlo simulation method. Four different sample sizes, n are used, namely $n = 50, 70, 150$ and 500 respectively. The data using the same procedure as described in Section 4.3 are generated. To assess the performance of the *COVRATIO* statistic, randomly introduce an outlier at a certain observation, for example, at the d^{th} observation. For the d^{th} observation, generate the data from the normal distribution with mean 0 and variance σ_d^2, where $\sigma_d^2 =$ 2, 4, 6, 8, 10 and 12. The data generated is then fitted by using the model in (2.2) and then the $\left|COV\right|$ is calculated using (4.1). Later on, the i^{th} row is excluded consequently from the sample, where $i = 1,...,n$. The reduced data is refitted by calculating $COVRATIO_{(-i)}$ using (4.2) and the maximum value of the $\left|COVRATIO_{(-i)}-1\right|$ is specified. The power of performance for $\left|COVRATIO_{(-i)}-1\right|$ statistic is calculated by computing the percentage of correct detection of the contaminated observation at the d^{th} observation. The R programming code for finding the power of performance for *COVRATIO* statistic can be found in Appendix D.

For illustration, Figure 4.10 shows the power of performance of $\left|COVRATIO_{(-i)}-1\right|$ statistic for $n = 50$ for $\sigma_\varepsilon =$ 0.2, 0.4, 0.6, 0.8, and 1.0. From this plot, it can be concluded that as σ_ε decreases, the power of performance in detecting the correct outlier increases. For other sample sizes of $n = 70,150,$ and 500, all these plots give similar results whereby, as σ_ε decreases, the power of performance in detecting the correct outlier increases. The results can be found in Appendix D.

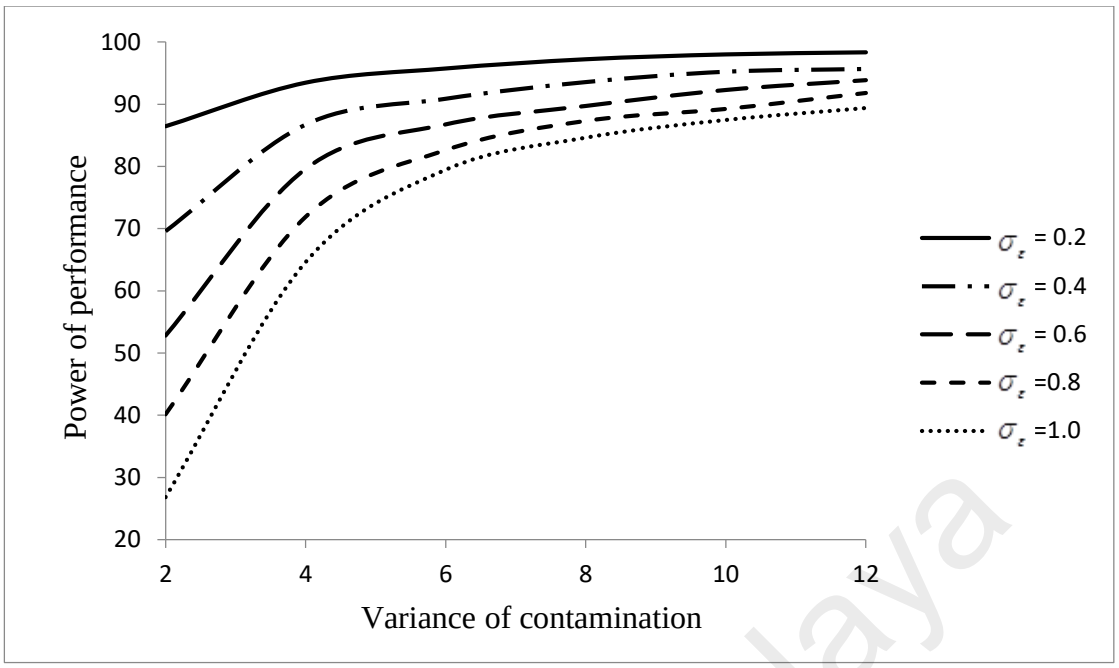

Figure 4.10: Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $n = 50$

The power of performance for a fixed σ_ε is investigated and the sample sizes, n are varied. Figure 4.11 shows the power of performance of $\left|COVRATIO_{(-i)} - 1\right|$ statistic for $\sigma_\varepsilon = 0.2$ for $n = 50, 70, 150$ and 500. From this result, it can be said that the power of performance is independent of the sample size. When $\sigma_\varepsilon = 0.4, 0.6, 0.8$ and 1.0, all these results gave consistent results whereby they are also independent of the sample size. Results can be found in Appendix D. In summary, the simulation study provides empirical evidence that as the variance of contamination increase, the power performance also increases.

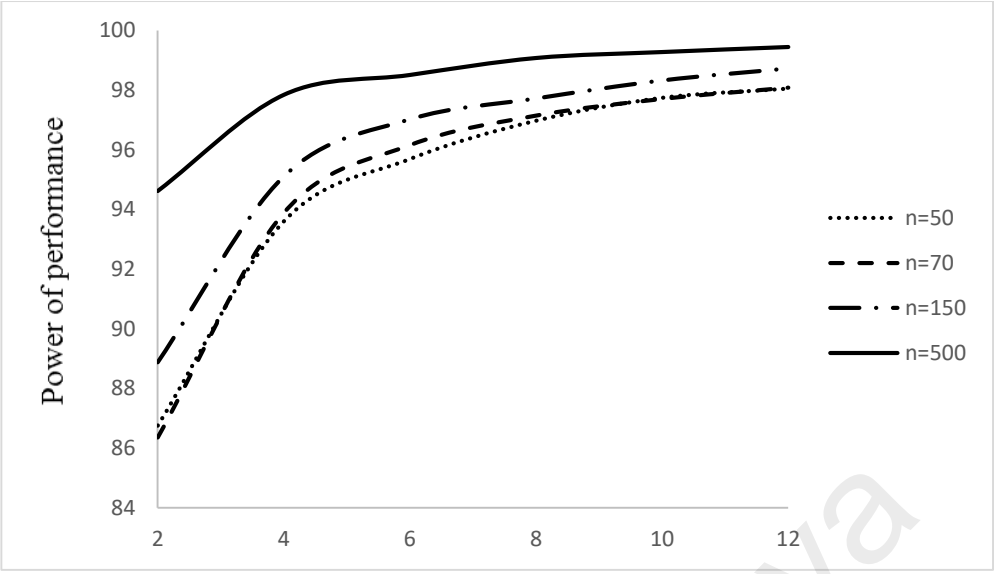

Figure 4.11: Power of performance for $\left|COVRATIO_{(-i)} - 1\right|$ when $\sigma_{\varepsilon} = 0.2$

4.5 Practical Example

For illustration, generate $n = 80$ data from LFRM, by setting the parameters $\alpha = 0, \beta = 1, \lambda = 1, \mu = 0,$ and $\sigma_{\delta}^2 = \sigma_{\varepsilon}^2 = 0.4^2$. The R Programming simulation code and the simulated data sets are presented in Appendix E, where the 20th observation of the data set is randomly contaminated by generating the contamination using $\varepsilon_i \sim N(0,16)$. The scatterplot of the generated data sets which includes the contaminated observation are presented in Figure 4.12. From the scatterplot, no outliers can be identified clearly. Even the 20th observation does not look like a candidate of an outlier.

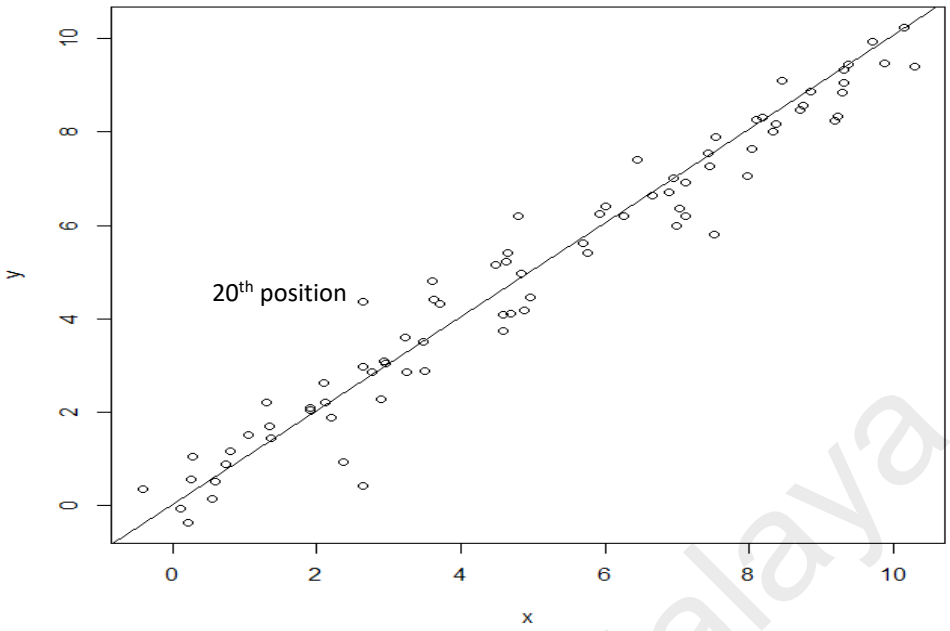

Figure 4.12: The scatter plot for the simulated data, n = 80

The *COVRATIO* statistic for each of the value is calculated and the results are given in Appendix F. Based on the formulation as given in Table 4.4, the cut-off point for $n = 80$ is calculated and the value 0.8538 is obtained as the cut-off point at 5% significant level. From Figure 4.13, it clearly shows that the 20th observations exceeds the cut-off points of 0.8538. To conclude, the developed test statistic is able to detect an outlier in the LFRM data set.

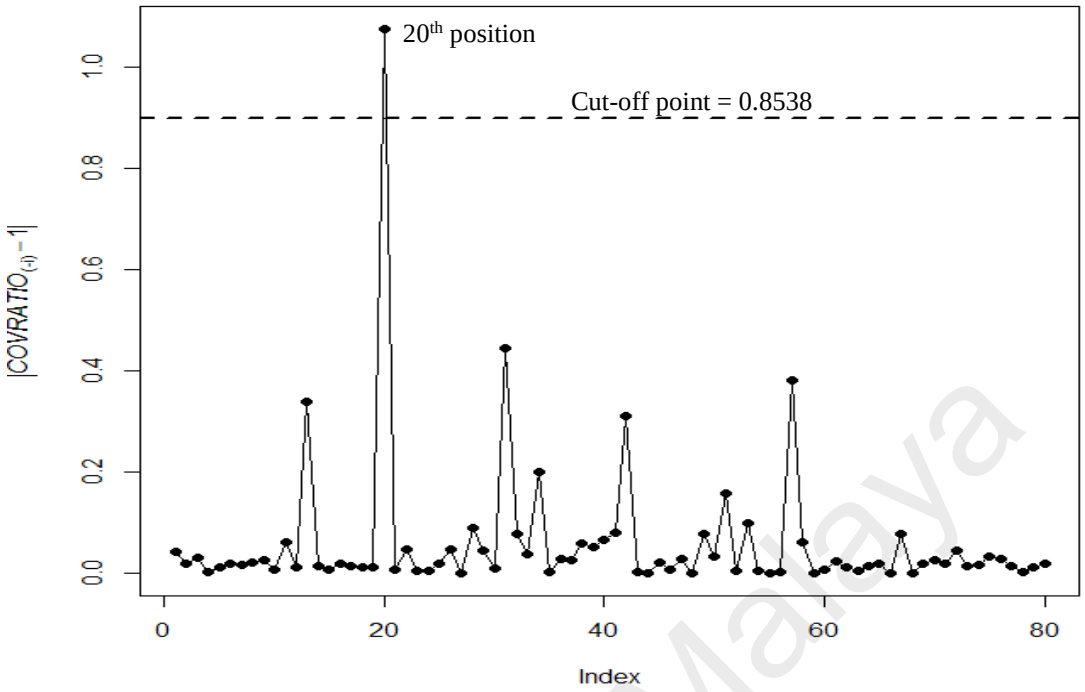

Figure 4.13: Graph of $\left|COVRATIO_{(-i)}-1\right|$ for simulation data, n = 80

4.6 Real Data Example

As another illustration, consider a real data set from a study conducted by Goran et al. (1996). The study was to examine the accuracy of some widely used body-composition techniques for children between the ages of 4 and 10 years by two different techniques, namely skinfold thickness (ST) and bioelectrical resistance (BR) and the data is given in Appendix A. The data fits in the LFRM, by having $\alpha = 0$ and $\beta = 1$ with additional assumption that the measurement error can occur in either variable of this experiment, as noted in Equation (2.1). The data are plotted in a scatterplot as shown in Figure 4.14. From the scatterplot, it is not easy to identify which observation is an outlier, if any. Even by looking at the 45th observation, it cannot be simply said that it is a candidate of an outlier, as other observation also seem to be quite far from the fitted line.

To confirm this, the $COVRATIO$ statistic is applied by plotting the $\left|COVRATIO_{(-i)} - 1\right|$ against the observation. From this plot, it can be seen that the 45th observation in Figure 4.15 has a value of $\left|COVRATIO_{(-i)} - 1\right|$ = 2.749. This value exceeds the cut-off point of 0.6855 for determining an outlier for sample size $n = 97$. Therefore, a conclusion can be made that the 45th observation is in fact, an outlier. Details of the R Programming to plot Figure 4.15 can be found in Appendix G.

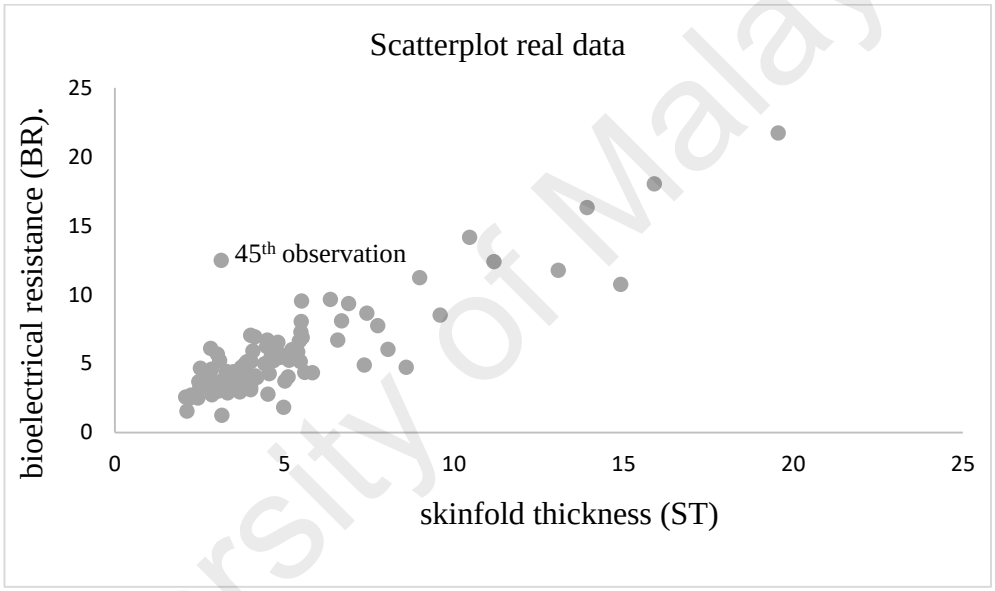

Figure 4.14: The Scatterplot for the real data, Skinfold Thickness (ST) and Bioelectrical Resistance (BR)

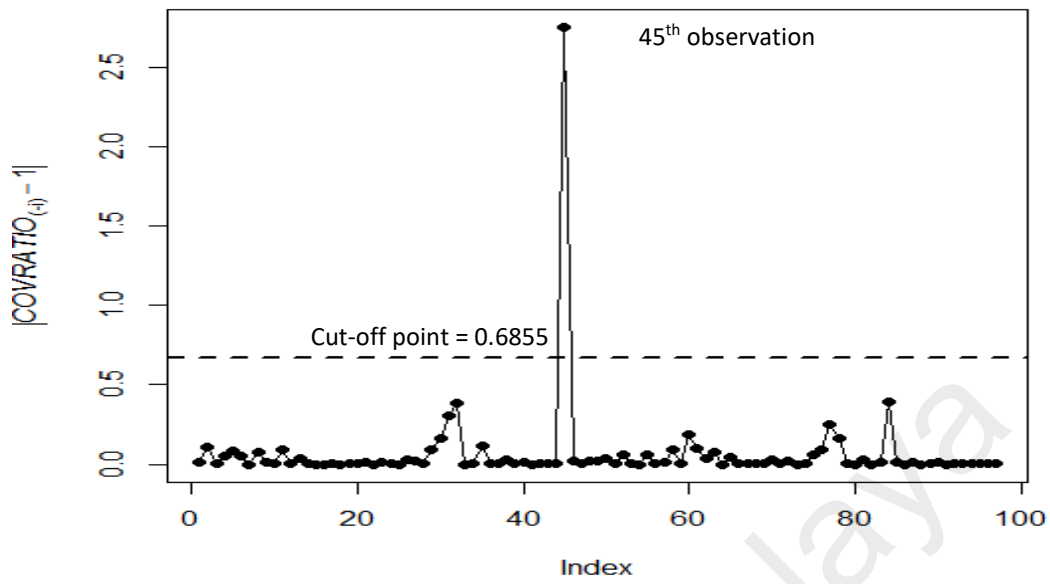

Figure 4.15: Graph of $\left|COVRATIO_{(-i)}-1\right|$ for real data with $n=97$.

After detecting the single outlier in Goran et. al (1996) data, the standard deviation (SD) of the parameters when outlier is present in the data and also when the outlier is excluded from the data is compared. The SD of the parameter estimation for both these two situation is given in Table 4.5. Looking at the SD of the parameters when no outlier exists in the data, the values of the SD of the parameters are much smaller compared to when there is an outlier in the dataset. This means that with the exclusion of the single outlier, the parameter estimates becomes more accurate. Therefore, identifying an outlier is necessary, so that the outlier observation can be excluded from the data, thus causing the parameter estimates to be more reliable.

Table 4.5: Parameter estimation and standard error of the estimated parameters

Goran et al. (1996) n=97, when outlier is included in the data			**Goran et al. (1996) n=96, when outlier is excluded from the data**		
Parameter estimation	Parameter estimation	Standard deviation	Parameter estimation	Parameter estimation	Standard deviation
$\hat{\alpha}$	0.0273	0.4610	$\hat{\alpha}$	0.0787	0.3287
$\hat{\beta}$	1.1285	0.0810	$\hat{\beta}$	1.0997	0.0572
$\hat{\sigma}_{\varepsilon}$	1.4400	1.2027	$\hat{\sigma}_{\varepsilon}$	1.0703	1.0346

4.7 Summary

In this chapter a test statistic based on *COVRATIO* is derived to detect a single outlier in the LFRM. A cut-off point can be expressed by the function $y = 321.04n^{-1.262}$ or 0.01 level of significant, $y = 135.63n^{-1.145}$ for 0.05 level of significant, and $y = 89.44n^{-1.090}$ for 0.10 level of significant respectively. Examining the power of performance for the cut-off points, it shows that the performance increases as the σ_{ε} decreases, for any n. Illustration using simulated and real data shows that the statistic works well in detecting a single outlier in LFRM.

CHAPTER 5: MULTIPLE OUTLIERS DETECTION IN LINEAR FUNCTIONAL RELATIONSHIP MODEL USING CLUSTERING TECHNIQUE

5.1 Introduction

In this chapter, an efficient technique to identify multiple outliers in linear functional relationship model by using the clustering technique is proposed. This work is an extension of the clustering algorithm as proposed by Sebert et al. (1998), which have been focused on the linear regression data. Section 5.2 describes in detail the similarity measure for identifying a cluster group for LFRM. This is followed by a single linkage clustering algorithm for LFRM in Section 5.3. Next, a robust stopping rule is proposed in clustering for LFRM as described in Section 5.4. In Section 5.5, a robust technique to identify the multiple outliers using a clustering method for LFRM is given. A simulation study in Section 5.6 is conducted to test the performance of this clustering algorithm in the LFRM. Finally, the usage of this clustering technique have been applied to classical data sets in Section 5.7. A summary is provided in the final section of this chapter.

5.2 Similarity Measure for LFRM

To group the variables or items into their own groups, it is necessary to have a certain measurement of similarity or a measure of dissimilarity between the items (Sebert et al., 1998). Therefore, finding the similarity measure is the first rule to cluster the items. There are four types of similarity measure which are correlation coefficient, distances measures, association coefficients and probabilistic similarity coefficients (Aldenderfer and Blashfield, 1984). All these four methods have its own strengths and drawbacks, so it is necessary to choose the best measurement that fits this model.

The most commonly used similarity measure by using the distance measure type is the Euclidean distance, defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}\ , \qquad (5.1)$$

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 (Wang et al., 2005). For the LFRM model, the Euclidean distance is used as the similarity measure because it can easily be applied, where by similar observations are identified by relatively small distance, while a dissimilar observation is identified by a relatively large distance. The following is an example of using the Euclidean distance as a similarity measure in LFRM, as summarized in Table 5.1.

Table 5.1: Observations x and y to illustrate Euclidean as a similarity measure

Observation	x	y
1	4.5525	4.2636
2	2.8234	6.0888
3	3.8888	5.1175
4	5.4915	8.0412
5	10.4554	14.1576

As an illustration, calculate the matrix of distance (similarity matrix) between all the possible pair of variables. The distance between observation 1 and 2, is calculated as

$$d_{12} = \sqrt{(4.5525 - 2.8234)^2 + (4.2636 - 6.0888)^2}$$

$$= \sqrt{2.9898 + 3.3314}$$

$$= 2.5142$$

Initially, since there are five observations in this illustration, place them in a square matrix with five rows and five columns. Then, the distance for $d_{ij} = d_{ji}$ is written in a similarity matrix in an upper triangular matrix. Table 5.2 shows the similarity matrix for this five observation.

Table 5.2: The similarity matrix for five observation

Observation	1	2	3	4	5
1	0	2.5142	**1.0815**	3.8926	11.5211
2		0	1.4417	3.3061	11.1064
3			0	3.3342	11.1733
4				0	7.8695
5					0

5.3 Single Linkage Clustering Algorithm for LFRM

After finding the suitable similarity measure for this model, the next step is to cluster the data. From Chapter 2, there are three major clustering techniques namely linkage, centroid, and Wards. In this study, the single linkage method is applied as the calculation is mathematically easy and has been widely used (Aldenderfer and Blashfield, 1984). Single linkage algorithm uses the smallest dissimilarity between a point in the first cluster and a point in the second cluster, and also defined as using the nearest neighbour (Kaufman and Rousseeuw, 1990).

The general steps for single linkage clustering algorithm in LFRM is explained in Figure 5.1. Find the smallest distance in $D = \{d_{ik}\}$, and merge the corresponding objects, say U and V, to get (UV). To calculate the distances between (UV) and other clusters, W as in Step 3 from Figure 5.1, compute the following,

$$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}, \tag{5.2}$$

where d_{UW} and d_{VW} are distances between the nearest neighbours of clusters U and W, and clusters V and W, respectively.

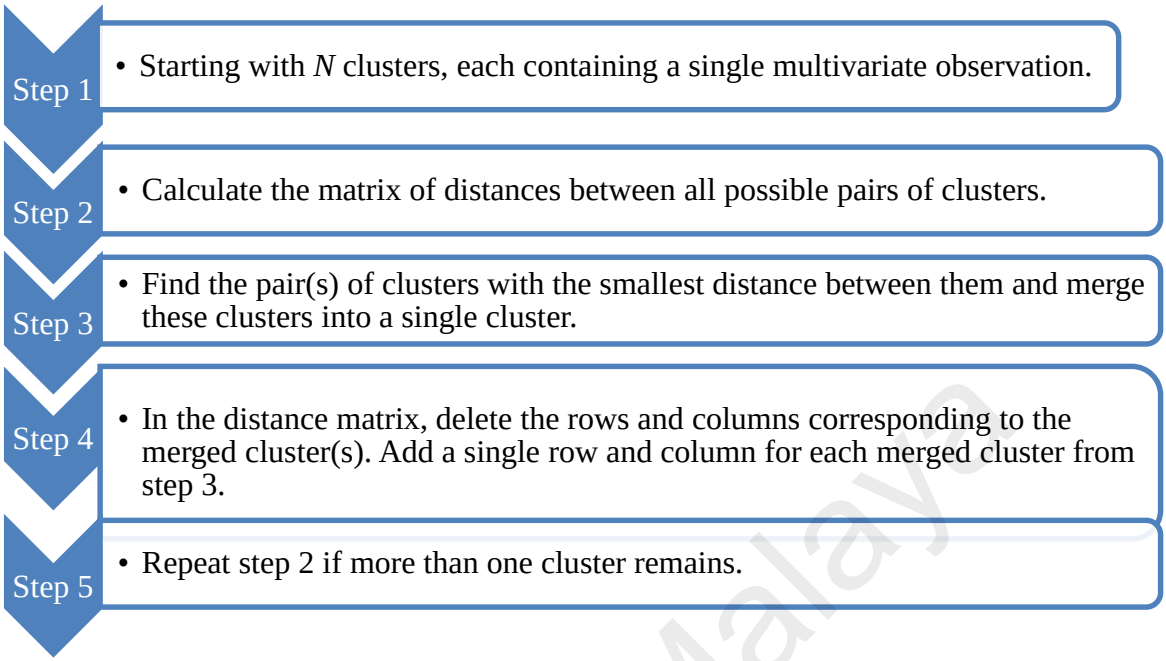

Figure 5.1: The general sequence in single linkage clustering algorithm

Here an example of the single linkage algorithm for a linear functional relationship model from five observations are illustrated in Table. 5.1. At the beginning, there will be five clusters, with one element at each cluster. Next, the distance (similarity matrix) is calculated, using Euclidean method as in (5.2) between all possible pairs of cluster. The matrix of distances for this particular illustration is given in Table 5.2.

Using a single linkage algorithm, merge the pair of clusters with the smallest distance first. From Table 5.2, the smallest distance is seen in cluster 1 and 3, which is 1.0815. In this step, cluster 1 and 3 is merged first, then the row 1 and column 3 in the similarity matrix are deleted. Table 5.3 shows the new similarity matrix with the new row and column which is added to cluster (1, 3).

Table 5.3: The new similarity matrix when (1, 3) is added

Observation	**2**	**4**	**5**	**(1,3)**
2	0	3.3061	11.1064	**1.4417**
4		0	7.8695	3.8926
5			0	11.5211
(1,3)				0

As the single linkage is based on the "nearest neighbours", the distance between cluster 2 and cluster (1, 3) is measured based on observation 2 and 3. Next, the distance between cluster 4 and cluster (1, 3) is measured based on observation 1 and 4. Meanwhile, the distance between cluster 5 and cluster (1, 3) is obtained based on observation 1 and 5. Later on, the corresponding column and row of cluster 2 and (1, 3) are deleted. Table 5.4 presents the similarity matrix with new row and column when cluster (2(1,3)) is added.

Table 5.4: The new similarity matrix when (2(1,3)) is added

Observation	**4**	**5**	**2(1,3)**
4	0	7.8695	**3.8926**
5		0	11.5211
2(1,3)			0

From Table 5.4, the shortest distance is between cluster 4 and (2,(1,3)), or known as cluster 4 and (1, 3), with the value 3.8926. Next, merge cluster 4 and (2(1,3)). Table 5.5 shows the similarity matrix with the new row and column when (4(2(1,3))) is added.

Table 5.5: The new similarity matrix when (4(2(1,3))) is added

Observation	**(4(2(1,3)))**	**5**
(4(2(1,3)))	0	**7.8695**
5		0

Finally, merge cluster 5 and cluster (4(2(1,3))) where all the observations are combined together in one cluster. The distance between cluster 5 and cluster (4(2(1,3))) is calculated using the distance between cluster 5 and 4, that is 7.8695. The results of the

single linkage clustering algorithm can be displayed in a form of a dendogram, or usually referred to as the "cluster tree" diagram. Figure 5.2 shows a general example of a cluster tree. The branches in the cluster tree represents clusters. The branches merge at nodes, positioning along a distance (or known as similarity) axis, that indicates the level at which the fusions take part, and in this example, the similarity axis is labelled as height.

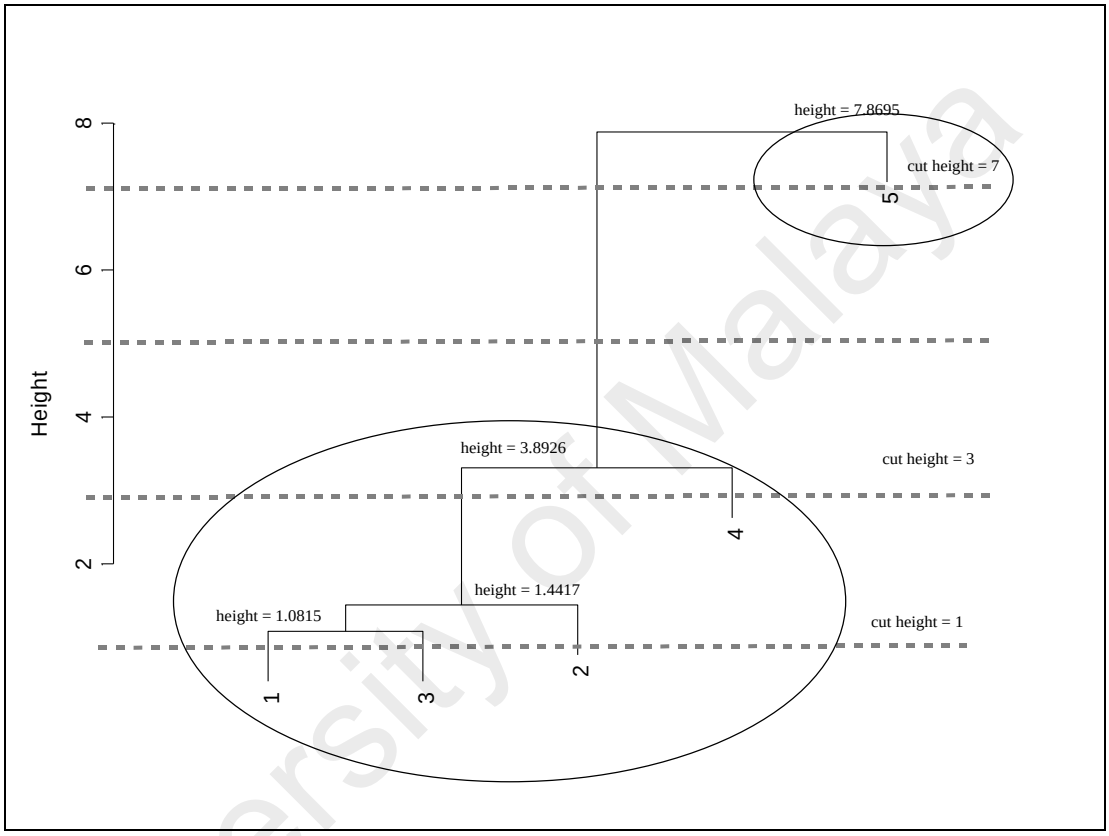

Figure 5.2: A general cluster tree for the single linkage algorithm

R Programming is used to illustrate the single linkage algorithm as a similarity measure in LFRM. The command line and output of the illustration is shown in Figure 5.3. The command line `d <- dist(data, method = "euclidean")` and `H.fit <- hclust(d, method="single")` shows that the similarity measurement used is the Euclidean distance, and the linkage method used is single linkage method.

```
data<-read.csv(file.choose(), header=TRUE)          # to call the data from excel
head(data)                 # the header of the column are the variables being used
d <- dist(data, method = "euclidean")                #similarity measurement
H.fit <- hclust(d, method="single")            # to plot the hierarchical cluster
h2<-H.fit
plot(h2)                                  # to display the dendogram of the clusters
```

Figure 5.3: The command in R programming for agglomerative hierarchical clustering

5.4 A Robust Stopping Rule for Outlier Detection in LFRM

After a cluster is obtained from the data, a user has to decide on the number of groups (if any) in the data set. The cluster tree needs to be portioned or "cut" at a certain height. As stated by Sebert et al. (1998), the number of cluster groups depend upon where the tree is cut. From the cluster tree in Figure 5.2, if the tree is cut at height equals to one, the data will be divided into three groups, and if the tree is cut at 7, the data will be divided into two groups.

Studies on stopping rule has been done by Milligan and Cooper (1985), but the difficulty is in a two clusters scenario which sees that a two-cluster case is the most difficult format to identify the stopping rules. Mojena's stopping rule on the other hand is widely used for linear variables (Mojena, 1977). Mojena's stopping rule, or known as "cut height" is $\bar{h} + \alpha s_h$, where $\bar{h}$ is the mean of heights for all $N-1$ clusters, and s_h is the unbiased standard deviation of the heights which is denoted in a specified constant. Mojena initially suggested that α should be specified in the range 2.75-3.50 as it gave the best overall results. However, Milligan and Cooper (1985) did an evaluation on 30

rules to determine the optimal number of clusters and concluded that the best overall performance of Mojena's stopping rule is when $\alpha = 1.25$.

The stopping rule proposed by Mojena (1977) uses the mean and standard deviation of the heights and this measurement may easily be affected in the presence of outliers (Hampel et al., 2011). In this chapter, a new stopping rule that will be more robust in the presence of outliers is proposed by using the relationship of the median and the median absolute deviation (MAD) of the tree heights. This method was introduced earlier by Midi (2010) to identify high leverage points in logistic regression model and they suggested the constant, c may be appropriately chosen as 2 or 3.

For this LFRM study, the following is proposed as a stopping rule to the clustering tree,

$$\bar{h} + cMAD(h), \tag{5.3}$$

where h are the cluster heights, $\bar{h}$ is the median of the heights for all $N-1$ clusters, $c = 3$ is the proposed constant variable for LFRM and $MAD(h)$ is the mean absolute deviation of the heights, defined by

$$MAD(h) = median|(h - median(h)|. \tag{5.4}$$

It can be said that with 95% confidence level that the cluster groups that exceeds this stopping rule will be classified as the potential outliers.

5.5 An Efficient Procedure to Detect Multiple Outliers in LFRM

To recap, in clustering multivariate data, an analyst must decide the point of variables to use, the measure of similarity to choose, and the clustering algorithm to use. Using the residual plotted against the corresponding predicted values is a good way to assess the model adequacy and is also a valuable tool to identify multiple outliers (Sebert et al., 1998). If there are no outliers present in the data, the observations will generally have a linear relationship that can be seen in the plot of the predicted and residual values.

In this section, an efficient clustering algorithm based on the single linkage method is proposed to cluster the points based on the predicted values and the residual values for the linear functional relationship model. The focus is to see how the clustering algorithm works, and how this clustering method is applicable to identify multiple outliers in LFRM. To summarize, the proposed algorithm consists of the following:

1) Variables to use: Predicted and residual values obtain from LFRM.
2) Measure of similarity to use: The Euclidean distance.
3) Clustering algorithm to use: Single linkage algorithm.
4) Stopping rule: The proposed robust cut tree as in (5.3).

Figure 5.4 presents the clustering algorithm that will be implement in LFRM. In this study, the Euclidean distance and the single linkage method is used to group the observation via clustering. The cluster that exceeds the proposed robust stopping rule will be identified as the potential outliers. Generally, the cluster groups with the largest observations are considered the clean observations, and all the other observations in a smaller cluster are considered as outliers (He et al., 2003). R programming for this algorithm is given in Appendix H. Next, the power of performance of the proposed

clustering technique in LFRM will be investigated via simulation study, which will be explained in Section 5.6.

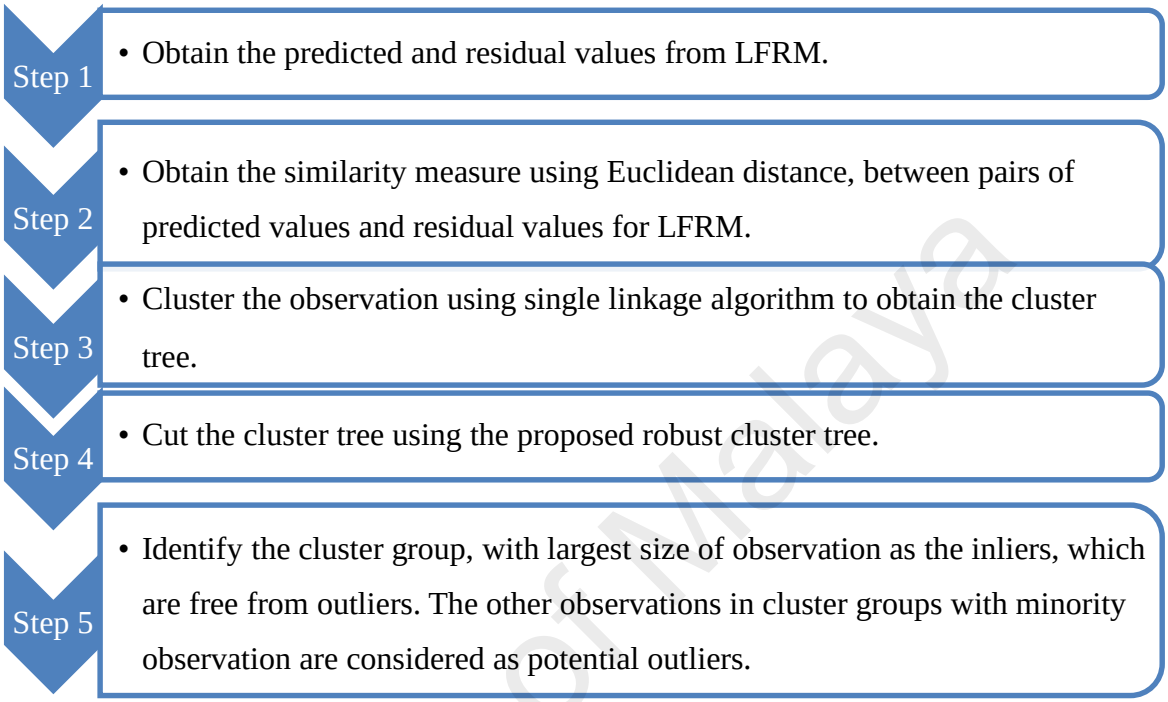

Figure 5.4: Flow chart of the steps in the proposed clustering algorithm for LFRM

5.6 Power of Performance for Clustering Algorithm in Linear Functional Relationship Model

There are two main issues that needs to be highlighted in identifying the multiple outliers, which are the masking and swamping affects. Masking occurs when an outlier is not detected, while swamping occurs when the inlying observation is mistakenly identified as an outlier. Masking may cause a more severe problem than swamping, as the inability to detect an influential observation can cause a dramatic influence to the model (Sebert et al., 1998).

Therefore, in this study, the performance of the proposed clustering algorithm in LFRM is investigated by measuring the “success” probability as well as the probability of swamping and masking obtained from the simulation study. Figure 5.5 shows how the performance of this clustering method is identified, whether it is swamping or masking.

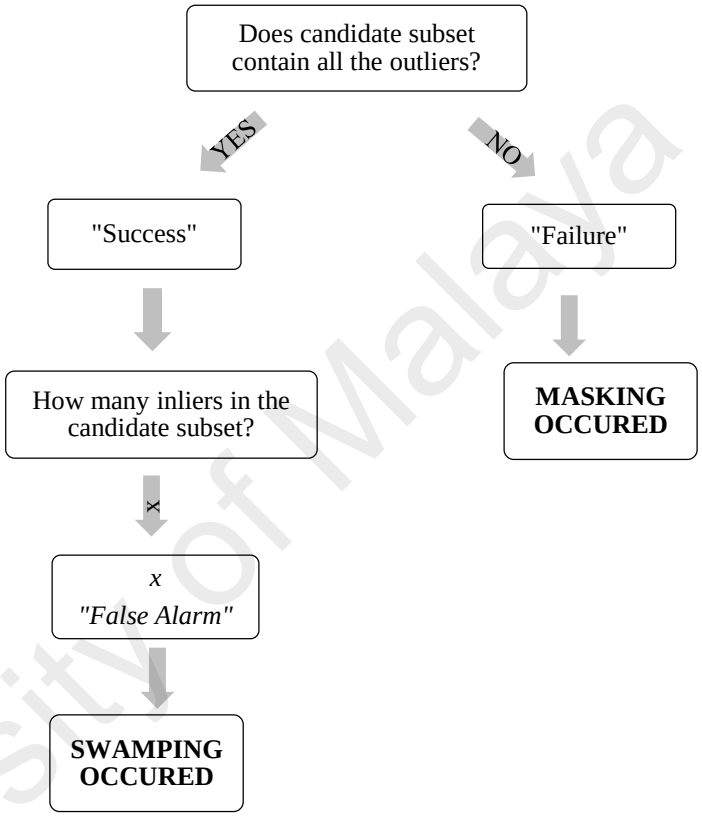

Figure 5.5: Flow chart of the clustering performances to check for swamping or masking cases

Assessment of the clustering method using this computation has been used by Sebert et al. (1998), Adnan and Mohamad (2003), and Satari (2014) respectively but for different models. “Success” defines the following clustering algorithm successfully identifies all the outlying observations, and no masking occurs in the observation. On the other hand, if the following clustering method is successful in identifying the outliers but

it also includes inlying observations as candidates of outliers (swamping occurs), then this will be identified as a "false alarm".

5.6.1 Simulation study

A simulation study is performed to assess how the level of contamination behave and to obtain the power of performance for the proposed clustering technique in LFRM. R Programming is used to perform the simulation study, as given in Appendix H. Random sample of sizes, $n = 50,70$ and 100 are generated respectively where the parameters are set to $\alpha = 1$, $\beta = 1$, $\sigma_\delta^2 = 0.1$, and $\lambda = 1$ respectively. The following equations would be,

$$Y_i = 1 + X_i, \quad x_i = X_i + \delta_i, \text{ and } y_i = Y_i + \varepsilon_i,$$

$$\text{where } X_i = 10\frac{i}{n} \text{ and } \delta_i, \varepsilon_i \sim N(0, 0.1) \tag{5.5}$$

From the generated sample, the predicted value, $\hat{X}_i$ and the residual value, $\hat{V}_i$ are calculated from the following equations;

$$\hat{X}_i = \frac{\lambda x_i + \hat{\beta}(y_i - \hat{\alpha})}{\lambda + \hat{\beta}^2} \text{ and } \hat{V}_i = y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right), \tag{5.6}$$

$$\text{where } \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + \left\{\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S_{xy}^2\right\}^{\frac{1}{2}}}{2S_{xy}},$$

and

$$\hat{\sigma}_\delta^2 = \frac{1}{(n-2)}\left\{\sum\left(x_i - \hat{X}_i\right)^2 + \frac{1}{\lambda}\sum\left(y_i - \hat{\alpha} - \hat{\beta}\hat{X}_i\right)^2\right\},$$

for

$$\bar{y} = \frac{1}{n}\sum y_i, \ \bar{x} = \frac{1}{n}\sum x_i, \ S_{xx} = \sum (x_i - \bar{x})^2,$$

$$S_{yy} = \sum (y_i - \bar{y})^2 \text{ and } S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}).$$

The errors δ_i and ε_i are generated by using $\delta_i, \varepsilon_i \sim N(0,0.1)$. In order to make some observation as outliers, randomly contaminate the observation by replacing the mean of the contamination, ε with 1, 2, 3, up until 10 respectively.

For example, at point $[d]$ of the response variable y, the observation $v[d]$ is contaminated as

$$v^*[d] = v[d] + \varepsilon, \tag{5.7}$$

where $v^*[d]$ is the contaminated observation at position $[d]$ and ε is the degree of contamination in the range of $1 \le \varepsilon \le 10$. With this, it allows the outlying observation to be placed away from the inliers. In this study, for each data set, randomly insert five outliers at certain points $[d_1, d_2, d_3, d_4, d_5]$. Then the clustering algorithm is used to identify these planted outliers for data sets $n = 50, 70,$ and 100 respectively. This simulation process is repeated 1000 times.

The power of performance of the proposed procedure is measured using the "success" probability (*pop*), masking error probability (*pmask*), and swamping error probability (*pswamp*). Let *s* be the total number of simulations and *out* is the number of planted outliers in the data set. Thus, the probability of planted outliers which are correctly detected (*pop*) is

$$pop = \frac{"success"}{s}, \tag{5.8}$$

where *"success"* is the number of data set that the method successfully identified all of the planted outlying observations. The probability of planted outliers that is falsely detected as inliers (*pmask*) is

$$pmask = \frac{"failure"}{(out)(s)}, \tag{5.9}$$

where *"failure"* is the number of outliers in the data set that is detected as inliers. Also, the probability of clean observations that is detected as outliers (*pswamp*) is

$$pswamp = \frac{"\ false"}{(n-out)s}, \tag{5.10}$$

where *"false"* is the number of inliers in the data set that are detected as outliers.

5.6.2 Results and Discussion for Simulation Study

The simulation results of the power of performance for the clustering technique in LFRM with $n = 50$ are shown in Table 5.6, that presents the power performance of the clustering method using the "success" probability (*pop*), the probability of masking (*pmask*) and also the probability of swamping (*pswamp*). From Table 5.6, for $n = 50,$ the probability of "success" increases as the mean of contamination, ε increases. As the contamination level reaches 5, the "success" probability shows the highest value of $pop = 1,$ and this value suggest a good performance. Looking at the value of *pmask,* as the level of contamination increases, the value of *pmask* decreases to a value of close to zero at $\varepsilon = 5.$ As for the *pswamp*, the value is also close to zero. A small value of *pmask* and *pswamp* is good as it shows that the clustering technique is reliable and is not affected by the fundamental problem usually seen in the clustering algorithm.

Table 5.6: The power of performance of the clustering method in LFRM using "success" probability (*pop*), probability of masking (*pmask*) and probability of swamping (*pswamp*) for $n = 50$.

Mean of contamination, ε	***Pop***	***Pswamp***	***pmask***
1	0.0570	0.0000	0.7366
2	0.5250	0.0000	0.2834
3	0.9510	0.0000	0.0162
4	0.9990	0.0000	0.0002
5	1.0000	0.0000	0.0000
6	1.0000	0.0000	0.0000
7	1.0000	0.0000	0.0000
8	1.0000	0.0000	0.0000
9	1.0000	0.0000	0.0000
10	1.0000	0.0000	0.0000

Alternatively, the results of the power of performance of the clustering method using the *pop*, *pmask*, and p*swamp* for $n = 50$ can be plotted in a graph, as shown in Figure 5.6. It can be seen that the *pop* increases as the mean of contamination, ε increases. From the *pswamp* graph, it can be observed that as the mean of contamination ε increases, the *pswamp* probability value decreases. Looking at the *pmask* graph, it is genereally consistent at the 0 value, even when the mean of contamination, ε increases.

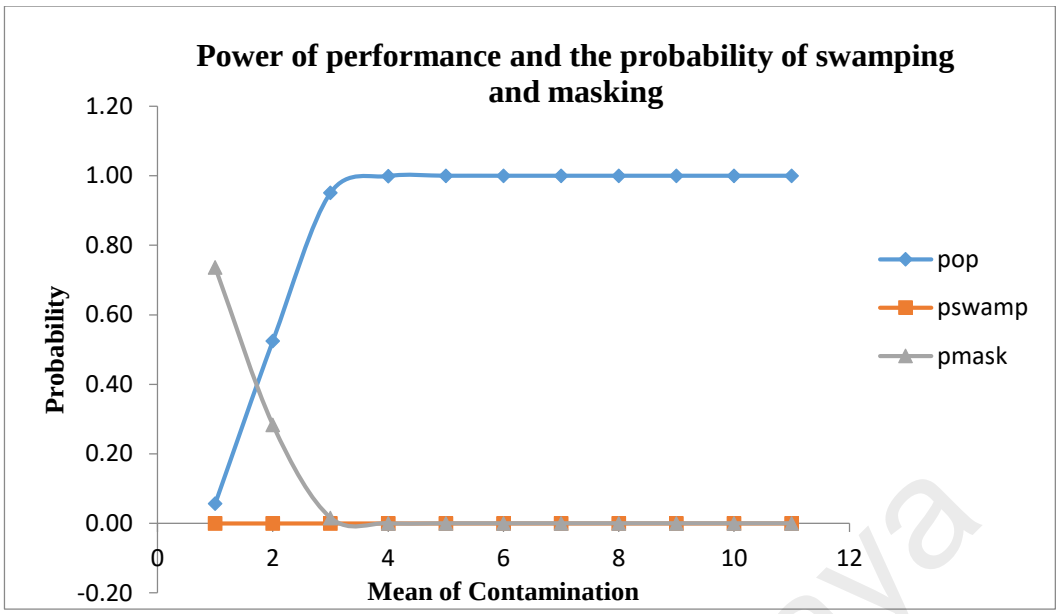

Figure 5.6: The plot of the "success" probability (*pop*), the probability of masking (*pmask*) and also the probability of swamping (p*swamp*) for $n = 50$.

Similar conclusions can be made when the power performance of sample size $n = 70$ and 100 are calculated. The results of the power of performance of the clustering method using the *pop*, *pmask* and *pswamp* for $n = 70$ and 100 are given in Appendix J and Appendix K.

In conclusion, from the "success" probability (*pop*), masking error (*pmask*) and swamping error (*pswamp)*, the proposed clustering method to identify multiple outliers in LFRM performs very well on simulated random data set. In general, at a higher level of contamination, ε the proposed method for LFRM gives a high value of *pop*, and a low value of *pmask* and *pswamp* and to summarize, the proposed clustering method performs the most efficient way if the outlying observation is located far from the remaining inlying observations.

5.7 Application to Real Data

As an illustration, two data sets are considered to demonstrate the applicability of the proposed clustering algorithm in a LFRM. The data sets are obtained from data sets that are used in many multiple outliers problem in linear regression model by Sebert et al. (1998). These data sets are usually referred to as "classic" multiple outlier data sets. Table 5.7 summarizes the methodology used by Sebert et al. (1998) in identifying the outliers in a linear regression model. In the table, p represents the number of regressor variables and n is the total number of observations in the data set. The outlying observation are the observations that are the potential outliers. The following two columns are the outlying observation that has been manage to identify using Sebert et al. (1998) methodology. The last column shows the number of observations that has been noted as swamping observation in the study.

Table 5.7: Sebert's et al. (1998) methodology performance on classical multiple outlier data sets

No	Data Sets	Outlying observation	Outlying observations identified	Number of observations swamped	Number of observations masked
1	Hertzsprung-Russell Stars Data (Rousseuw and Leroy, 1987)	11,20,30 and 34	11,20,30,34,7 and 14	2	0
2	Telephone Data (Rousseuw and Leroy, 1987)	15-24	15-24	0	0

First, the Hertzsprung-Russell Stars Data is used, and assuming measurement errors can occur at both variables, the proposed clustering algorithm in LFRM is applied. The x and y variables are plotted in a scatterplot as shown in Figure 5.7. From the

scatterplot, there are four observations that seems to be lying away from the other observations, which are observation 11, 20, 30 and 34. To correctly identify whether they are the outlying observation, the clustering process is applied in the LFRM and the proposed robust stopping rule to cut the tree.

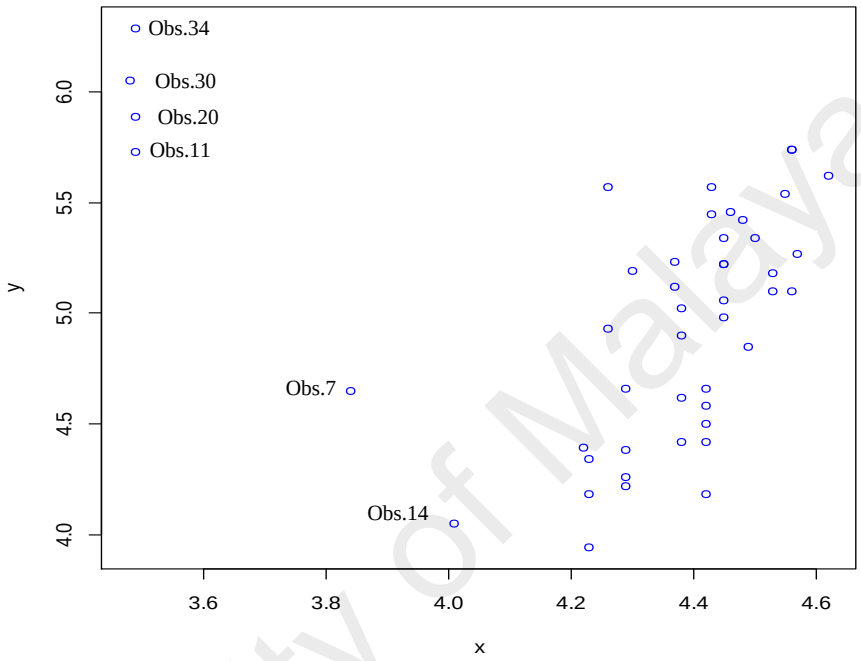

Figure 5.7: The scatterplot of Hertzsprung-Russell Stars Data

A clustering algorithm is done by clustering the predicted, $\hat{X}_i$ and residual, $\hat{V}_i$ values from the LFRM using the single linkage algorithm, and the Euclidean distance as the similarity measure. Based on the proposed robust stopping rule as in (5.3), the tree will be cut at $\bar{h} + 3MAD(h) = 0.4903$. From the R Programming result, as given in Appendix I, two clusters are formed, one cluster containing the majority of the observation, and another smaller cluster containing observation 7, 11, 14, 20, 30, and 34 as shown in Figure 5.8. It can be seen that the proposed clustering technique for the data in the LFRM successfully identified outliers in observation 11, 20, 30 and 40. Observations 7 and 14 have been detected as the swamping observation in this study.

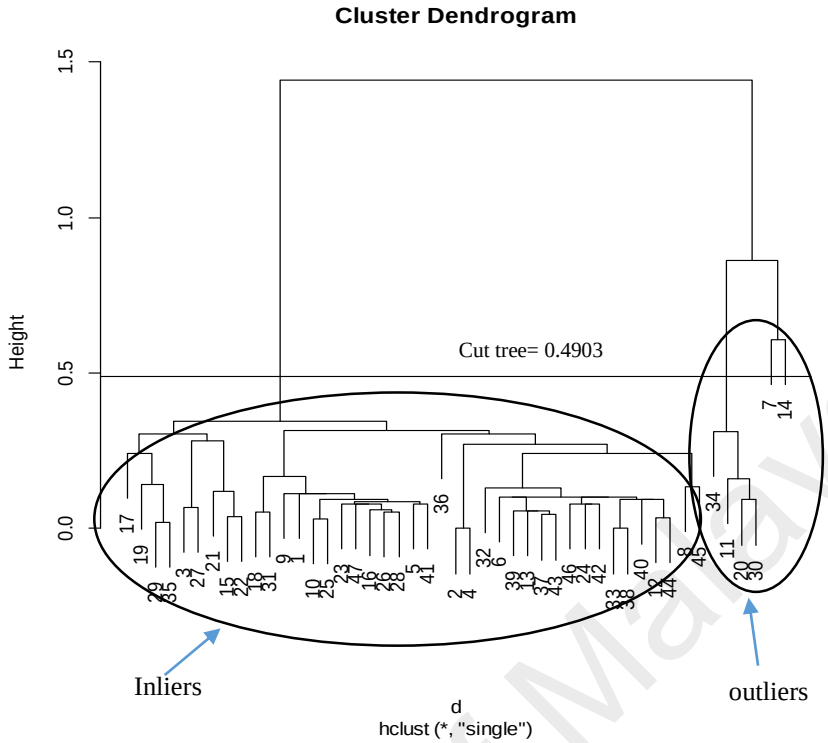

Figure 5.8: The cluster tree for Hertzsprung-Russell Stars Data

Next, the proposed clustering procedure for LFRM is applied to the Telephone Data by Rousseuw and Leroy, 1987. The scatterplot of x and y variables are shown in Figure 5.9. From the graph, observation 15 till observation 24 seems to be lying away from the other observation. Next, the clustering process in LFRM is applied and the proposed robust stopping rule at 1.4398 is obtained, as shown in Figure 5.10. It can be seen that the cluster tree is cut, leaving three clusters. One cluster contains the majority of the observations, and another two smaller clusters containing the outlying observations, which are observation 15 till observation 24. It can be seen that that the proposed clustering technique for the data in the LFRM successfully identified all the outliers in the classic Telephone Data.

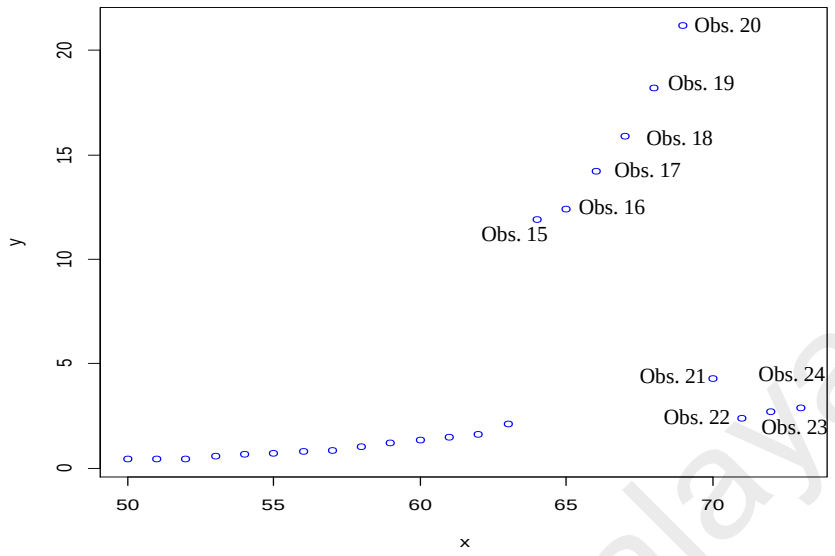

Figure 5.9: The scatterplot for Telephone Data

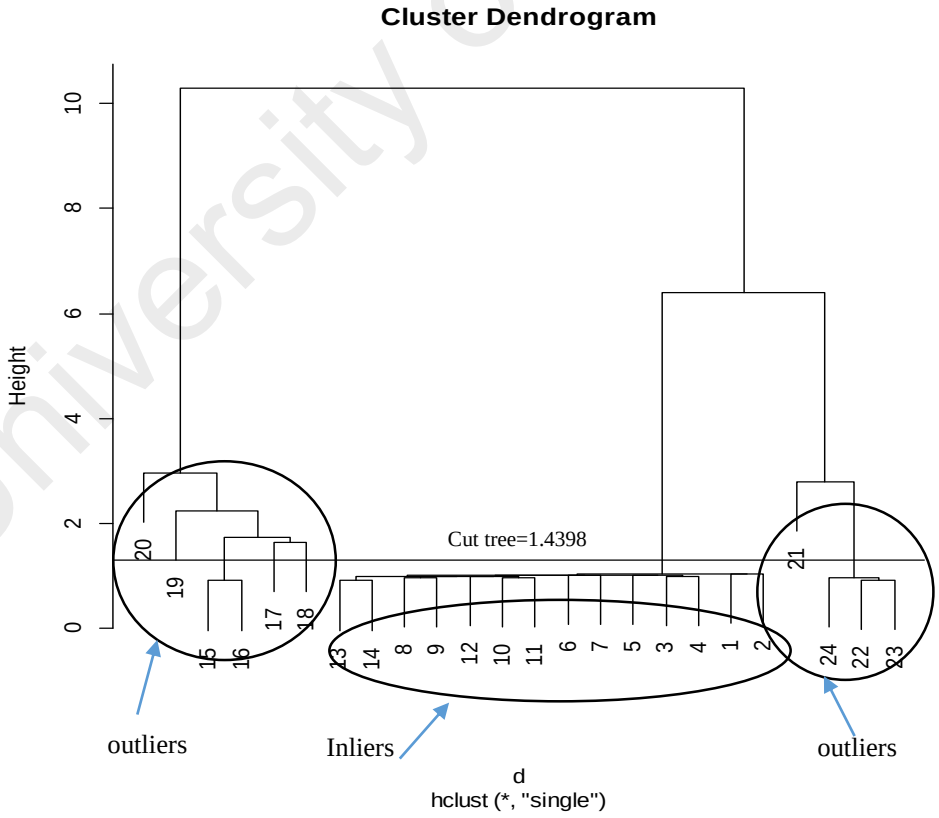

Figure 5.10: The Cluster tree for Telephone Data

5.8 Summary

An efficient procedure to identify multiple outliers in the LFRM is proposed using the single linkage algorithm and Euclidean distance as the similarity measure as proposed by Sebert et al. (1998) in a regression model. In this study, a robust cut tree using the relationship of the median and the median absolute deviation (MAD) of the tree heights is proposed and say that with 95% confidence level that the cluster group that exceeds this proposed cut tree, $\bar{h} + 3MAD(h)$ will be classified as the potential outliers.

From the simulation study, this proposed method is able to identify the planted multiple outliers in different sample sizes, n and with different mean of contamination on the outliers, ε. The probability of swamping and masking is practically small and at certain level of contamination, it becomes zero and this is good as the two main issues in multiple outlier detection is able to solve. Application in real data also shows that this proposed clustering method for the LFRM successfully detects the outliers as found in other classical data sets.

CHAPTER 6: MISSING VALUE ESTIMATION METHODS IN LINEAR FUNCTIONAL RELATIONSHIP MODEL

6.1 Introduction

This chapter reviews the missing value estimation methods for data that are in the LFRM. Section 6.2 describes the expectation-maximization (EM) algorithm and the expectation-maximization with bootstrap (EMB) algorithm as a modern imputation technique to handle missing values. Section 6.3 investigates the applicability of the EM and EMB methods in dealing with missing values for two types of LFRM. Section 6.4 measures the performance of the imputation method by using the EM and the EMB algorithm. A simulation study is performed to investigate the performance of EM and EMB in Section 6.5, while Section 6.6 illustrates the application of EM and EMB in real data example. Summary of the chapter is given in Section 6.7.

6.2 Imputation Methods

In the literature review chapter, the traditional and the modern techniques to handle missing value problems is discussed. Imputation methods are the most commonly used method to solve missing data (Little & Rubin, 2014). Traditional imputation methods include mean imputation, hot-deck imputation, and stochastic imputation (George et al., 2015). On the other hand, the modern imputation approaches include those based on maximum likelihood and multiple imputations (Acock, 2005). EM algorithm is an example of maximum likelihood and some examples of multiple imputations include Markov Chain Monte Carlo, Fully Conditional Specification, and EMB algorithm (Baraldi & Enders, 2010; Barzi & Woodward, 2004; Gold & Bentler, 2000; Little & Rubin, 2014). Modern approaches are favourable over the traditional approaches as they

require less assumption, they give less biased estimates, and they are advantageous as the data are not "thrown away" (Baraldi & Enders, 2010). In this chapter, two modern imputation techniques are compared, which are the expectation-maximization algorithm and the expectation-maximization with bootstrapping algorithm, which will be abbreviated as EM and EMB, respectively.

6.2.1 Expectation-Maximization Algorithm (EM)

EM algorithm is one example of an imputation method using the maximum likelihood, where it finds the maximum likelihood estimates through an iterative algorithm when there are missing values in the dataset (Little & Rubin, 2014). In short, EM will "fill in" the Y_{mis}, which are the missing data, based on an initial estimate of θ (where by the estimate of θ is found by using only the data that are observed). Then, θ is re-estimated based on Y_{obs} , which are the data that are observed, and the filled-in Y_{mis}, and this process is iterated until the estimates converge (Howell, 2008).

Figure 6.1 describes the EM process. To elaborate, EM comprises of two steps namely the expectation or E-Step, and the maximization or M-Step. In the E-step, the missing values are imputed by replacing Y_{mis} with the expected value of $E(Y_{mis} \mid Y_{obs}, \theta)$, by assuming $\theta = \theta^{(t)}$. Next, in the M-step, the expected value that is obtained from E-step will be maximised. These two steps will be done iteratively until it converges to a local maximum of the likelihood function (Schafer, 1997). A detailed explanation on the convergence properties of EM algorithm can be found in some literature, as an example by Wu, 1983.

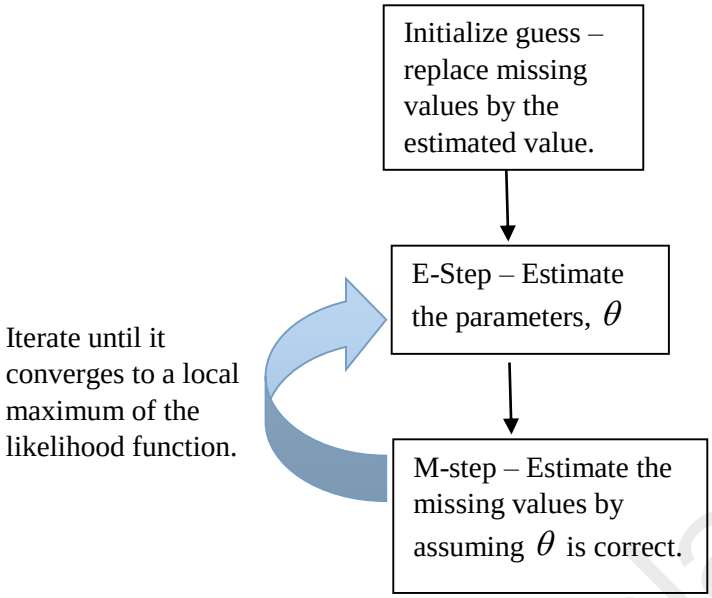

Figure 6.1: Flow chart of the expectation-maximization (EM) process

EM algorithm has become popular because of its simplicity, the generality of its theory and because of its wide application (Dempster et al. 1977). Several examples of the applicability of the EM include handling missing data in air pollutants studies (Schafer 1997), in survival model (Wang & Miao 2009) and in the linear regression model (Junger & de Leon 2015).

6.2.2 Expectation-Maximization with Bootstrapping (EMB) Algorithm

The emerging EMB algorithm is similar to the regular EM algorithm. However, it involves multiple nonparametric bootstrap samples of the original incomplete data. Figure 6.2 explains this algorithm in detail. The EMB algorithm performs multiple imputations that "fills in" the missing values in the incomplete data set. Multiple imputations are less biased and its efficiency is higher than the listwise deletion (Honaker et al., 2013; Rancoita et al., 2015).

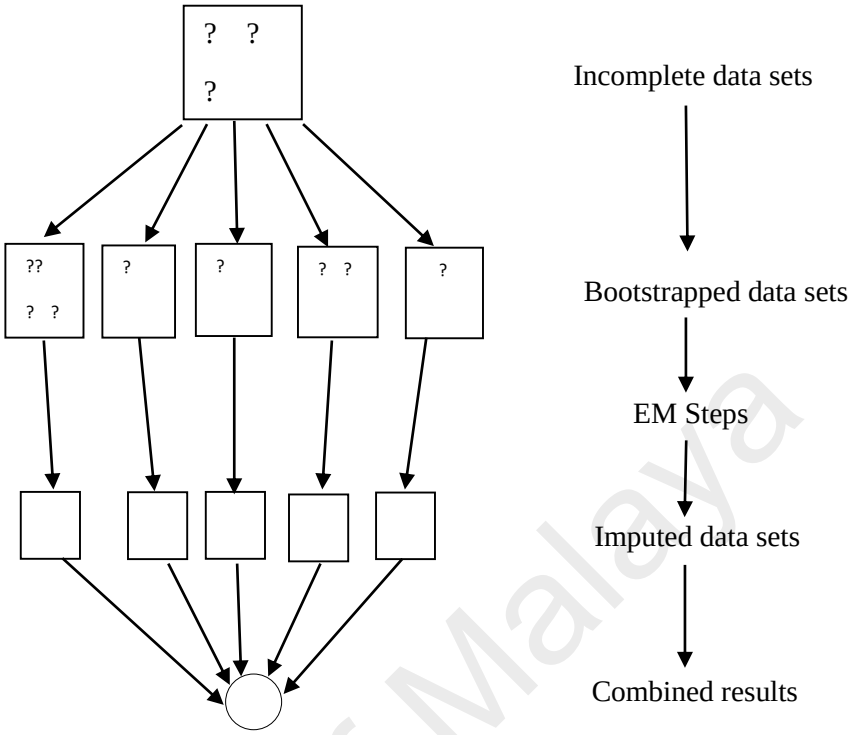

Figure 6.2: Multiple imputation using Expectation-maximization with bootstrap (EMB) algorithm

Applying multiple imputations can be challenging as the nature of its algorithm can be quite complicated, but with the available of high performance computing, it can help perform the multiple imputations in a much advanced way (Honaker et al. 2013).

6.3 Application of the EM and EMB method in Linear Functional Relationship Model

Studies on handling missing values using EM have been largely explored for the univariate data (Schafer, 1997), the linear regression data (Faria and Soromenho, 2010), the survival data, and the linear structural model (Mamun et al., 2016). The application of EM in handling missing data in LFRM, however, has not been explored. In this chapter, the EM and EMB methods in dealing with missing values for a type of model called the LRFM is proposed. A LRFM is often employed when the objective of the study is to compare two sets of data with both observable errors. The parameter estimates of LRFM can be obtained by the maximum likelihood estimates, which is referred as the full model LFRM with the acronym LFRM1, and when the slope parameter of LFRM is estimated using nonparametric approach, which is referred with the acronym LFRM2.

6.3.1 Linear Functional Relationship Model for Full Model (LFRM1)

As mentioned earlier, LFRM can be expressed by $Y = \alpha + \beta X$ where both variables X and Y are linearly related but observed with error. Parameter α is the intercept value, and β is the slope parameter. For any fixed X_i, the x_i and y_i are observed from continuous linear variable subject to errors δ_i and ε_i respectively, as given in (2.2). The error terms δ_i and ε_i are assumed to be mutually independent and normally distributed random variables, as seen in (2.3). There are $(n+3)$ parameters to be estimated, namely $\alpha, \beta, \sigma_\delta^2$ and $X_1,...,X_n$, the incidental parameters respectively and these parameters can be obtained from (2.5). Here, as the $\hat{\beta}$ is estimated using the maximum likelihood estimation method, it will be denoted as the LFRM1.

6.3.2 Linear Functional Relationship Model with nonparametric slope parameter estimation (LFRM2)

Alternatively, the parameter $\hat{\beta}$ can be obtained using nonparametric estimation as explained in Chapter 3. Hence, by assuming known slope as given by the $\hat{\beta}_{new}$ in (3.4), a robust estimate of the parameters in the linear functional relationship model is denoted by LFRM2.

6.4 Performance Measurement of EM and EMB

In order to measure the performance of the imputation using EM and EMB algorithm, several measurements are used, namely the mean absolute error (MAE), the root-mean-square error (RMSE) and the estimated bias (EB). MAE is the average of the difference between the predicted and actual data points (Junninen et al., 2004) and is given by

$$MAE = \frac{1}{N}\sum |P_i - O_i| \ , \tag{6.1}$$

where N is the number of imputations, P_i are the imputed values, and O_i are the observed data values. Values of MAE can be from 0 to infinity in which a value of zero is an indicative of a perfect fit.

RMSE measures the differences between the predicted and actual data points, and is given by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[P_i - O_i\right]^2} \ , \tag{6.2}$$

with N is the number of imputations, P_i and O_i are the imputed and observed data points, respectively (Lindley 1947). A small value of RMSE suggests a good fit and large value otherwise.

Mean of estimated biased (EB) of a parameter on the other hand is defined by

$$EB = mean\left|O_i - E_i\right|, \tag{6.3}$$

where the mean of the absolute difference of O_i, the estimated value of the parameters obtained from the observed data and E_i, the estimated value of the parameters obtained from the data after imputing the missing values is calculated. A small EB is indicative of a reliable performance estimator (Lindley, 1947).

6.5 Simulation Study

A simulation study is conducted to investigate the performance of these two imputation methods namely the EM and the EMB method. For the first simulation study, the LFRM1 is used as in 2.1, where without any loss of generality, the parameters are set to $\alpha = 1,\ \beta = 1,\ \sigma_\delta^2 = 0.1,$ and $\lambda = 1$, with sample sizes, $n = 50$ and 100 respectively. For the simulation study, the missing data are assumed to be missing at random (MAR), and are inserted randomly at 5%, 10%, 20% and 30% levels respectively (Howell, 2008). This simulation is conducted for 5000 trials, and the MAE, RMSE and EB of these two imputation methods, namely EM and EMB are analysed.

From Table 6.1 and Table 6.2, it can be observed that both methods perform well, with small MAE and small RMSE at each *n.* The EMB algorithm has significantly smaller MAE and RMSE values compared to the EM method for $n = 50$ and 100 respectively. For each level of percentage missing namely at 5%, 10%, 20% and 30% respectively, the EMB algorithm consistently gives smaller MAE and RMSE values as compared to the

the EM algorithm. Looking at the RMSE for $n = 50,$ at 5% missing values, the EMB algorithm show values which are different at only two decimal points from the EM algorithm. It is worthwhile to note that the percentage change from the EM to the EMB shows a significant difference of about 8.99% of improvement in the RMSE values. Another example, for the 20% missing value at $n = 50,$ the difference is significant with 15.23% improvement from the EM to the EMB algorithm. This proves that even though the difference of RSME is at two decimal places, it shows a huge improvement of the EM to the EMB algorithm. Note that as the sample size increase from $n = 50$ to $n = 100$, the RMSE values of the EMB decrease at all levels of percentage of missingness. This suggest that at a higher n, it leads to a smaller RMSE and a smaller bias.

Table 6.1: MAE and RMSE for LFRM1 using two imputation methods for $n = 50$

Percentage of missing (%)	Performance Indicator / Method	MAE	Percentage change of MAE (%)	RMSE	Percentage change of RMSE (%)
5%	EM	3.7530	25.13	5.4943	8.99
	EMB	2.8100		5.0003	
10%	EM	6.2616	19.82	5.1894	6.62
	EMB	5.0210		4.8457	
20%	EM	5.3612	17.85	4.9344	15.23
	EMB	4.4042		4.1827	
30%	EM	5.2312	13.21	5.4744	11.31
	EMB	4.5404		4.8550	

Table 6.2: MAE and RMSE for LFRM1 using two imputation methods for $n = 100$

Percentage of missing (%)	Performance Indicator / Method	MAE	Percentage change of MAE (%)	RMSE	Percentage change of RMSE (%)
5%	EM	4.6860	30.04	5.6699	32.43
	EMB	3.2781		3.8314	
10%	EM	5.2109	31.64	4.7434	16.36
	EMB	3.5623		3.9672	
20%	EM	5.6734	49.06	6.1135	26.60
	EMB	2.8900		4.4872	
30%	EM	4.4952	25.22	5.5477	17.99
	EMB	3.3617		4.5497	

From Table 6.3 and Table 6.4, it can be observed that using the EM and the EMB algorithm, both methods give small value for the mean of the estimated bias for all the parameters $\alpha, \beta,$ and σ_δ^2. Imputation using EMB method, however gives better precision with consistently even smaller bias values for all parameters as compared to the EM method. Looking at the standard error of each parameter in the parenthesis, it shows that at each level of missingness, the EMB outperforms the EM by having smaller values of standard error. These observations clearly indicate the superiority of the EMB method in comparison to the EM method.

Table 6.3: Mean of estimated bias and (standard error) of the parameters for LFRM1 using two imputation methods for $n = 50$

Percentage of missing (%)	Parameters/ Methods	α (standard error)	β (standard error)	σ_δ^2 (standard error)
5%	EM	3.621E-02 (3.001E-02)	6.600E-03 (5.321E-03)	5.101E-04 (4.231E-04)
	EMB	3.024E-02 (1.503E-02)	6.071E-03 (2.952E-03)	4.403E-04 (3.395E-04)
10%	EM	2.986E-02 (2.228E-02)	5.643E-03 (4.225E-03)	8.028E-04 (7.220E-04)
	EMB	2.865E-02 (2.208E-02)	5.510E-03 (4.217E-03)	6.829E-04 (5.599E-04)
20%	EM	3.086E-02 (2.291E-02)	5.613E-03 (4.193E-03)	1.147E-03 (1.010E-03)
	EMB	2.939E-02 (2.176E-02)	5.496E-03 (4.076E-03)	9.250E-04 (7.372E-04)
30%	EM	2.144E-02 (1.619E-02)	3.915E-03 (2.942E-03)	7.425E-04 (6.027E-04)
	EMB	2.079E-02 (1.552E-02)	3.895E-03 (2.909E-03)	5.904E-04 (4.477E-04)

Table 6.4: Mean of estimated bias and (standard error) of the parameters for LFRM1 using two imputation methods for $n = 100$

Percentage of missing (%)	Parameters/ Methods	α (standard error)	β (standard error)	σ_δ^2 (standard error)
5%	EM	2.012E-02 (1.538E-02)	3.909E-03 (2.983E-03)	3.837E-04 (3.437E-04)
	EMB	2.000E-02 (1.485E-02)	3.907E-03 (2.899E-03)	3.157E-04 (2.456E-04)
10%	EM	2.064E-02 (1.545E-02)	3.944E-03 (2.960E-03)	5.340E-04 (4.489E-04)
	EMB	1.989E-02 (1.514E-02)	3.827E-03 (2.894E-03)	4.301E-04 (3.375E-04)
20%	EM	2.132E-02 (1.616E-02)	3.892E-03 (2.974E-03)	7.952E-04 (6.492E-04)
	EMB	2.053E-02 (1.567E-02)	3.847E-03 (2.915E-03)	6.271E-04 (4.905E-04)
30%	EM	2.244E-02 (1.675E-02)	3.923E-03 (2.956E-03)	9.732E-04 (7.761E-04)
	EMB	2.093E-02 (1.612E-02)	3.835E-03 (2.921E-03)	7.636E-04 (5.913E-04)

The study is also replicated for the LFRM2, in which the slope parameter β is estimated using a nonparametric method. From the results as presented in Table 6.5 and Table 6.6, both methods of imputations are good, but EMB algorithm shows consistently smaller MAE and RMSE as compared to the EM algorithm for both $n = 50$ and 100. Note that, as the percentage of missing data increases, the EMB outperforms the EM in

terms of smaller MAE and RMSE values. Similar to LFRM1, the RMSE values of EM and EMB differs at only two decimal places but by looking at the percentage of improvement from EM to EMB, the change is significant. Again, the superiority of EMB applies for the LFRM2. Likewise, as n increases from 50 to 100, both MAE and RMSE suggest a better precision for the LFRM2.

Table 6.5: MAE and RMSE for the LFRM2 by using two imputation methods for $n = 50$

Percentage of missing (%)	Performance Indicator / Method	MAE	Percentage change of MAE (%)	RMSE	Percentage change of RMSE (%)
5%	EM	7.3911	27.67	9.0257	39.02
	EMB	5.3460		5.5041	
10%	EM	5.6922	48.60	7.3679	15.72
	EMB	2.9257		6.2096	
20%	EM	5.3877	4.52	5.7500	7.44
	EMB	5.1443		5.3224	
30%	EM	3.4405	6.86	4.8878	10.43
	EMB	3.2045		4.3782	

Table 6.6: MAE and RMSE for the LFRM2 by using two imputation methods for $n = 100$

Percentage of missing (%)	Performance / Indicator / Method	MAE	Percentage change of MAE (%)	RMSE	Percentage change of RMSE (%)
5%	EM	6.0935	36.73	5.4978	2.60
	EMB	3.8556		5.3549	
10%	EM	5.1017	53.49	5.2083	-0.67
	EMB	2.3729		5.2433	
20%	EM	3.7023	6.07	5.4531	25.83
	EMB	3.4775		4.0445	
30%	EM	3.9048	18.18	4.3791	4.73
	EMB	3.1950		4.1721	

For the measure of estimated bias as given in Tables 6.7 and 6.8, both methods give small values for the mean of the estimated bias for all the parameters $\alpha, \beta,$ and σ_δ^2. Imputation using the EMB method, however gives better precision with smaller bias values for all parameters as compared to the EM method. From the standard error of each parameter in the parenthesis, it shows that at each level of missing data, the EMB method outperforms the EM method by having smaller values of standard error. These observations clearly indicate the superiority of the EMB algorithm in comparison to the EM algorithm.

`

Table 6.7: Mean of estimated bias and (standard error) of the parameters for LFRM2 using two imputation methods for n =50

Percentage of missing (%)	Parameters/ Methods	α (standard error)	β (standard error)	σ_δ^2 (standard error)
5%	EM	3.069E-02 (2.329E-02)	5.944E-03 (4.511E-03)	4.899E-04 (4.812E-04)
	EMB	3.032E-02 (2.280E-02)	5.915E-03 (4.451E-03)	4.320E-04 (3.857E-04)
10%	EM	3.125E-02 (2.407E-02)	5.930E-03 (4.588E-03)	7.795E-04 (7.035E-04)
	EMB	3.027E-02 (2.294E-02)	5.828E-03 (4.432E-03)	6.445E-04 (5.249E-04)
20%	EM	3.258E-02 (2.436E-02)	6.007E-03 (4.523E-03)	1.153E-03 (9.966E-04)
	EMB	3.169E-02 (2.404E-02)	5.976E-03 (4.489E-03)	9.235E-04 (7.311E-04)
30%	EM	3.390E-02 (2.543E-02)	5.968E-03 (4.511E-03)	1.449E-03 (1.225E-03)
	EMB	3.292E-02 (2.451E-02)	6.081E-03 (4.498E-03)	1.190E-03 (9.659E-04)

Table 6.8: Mean of estimated bias and (standard error) of the parameters for LFRM2 using two imputation methods for n =100

Percentage of missing (%)	Parameters/ Methods	α (standard error)	β (standard error)	σ_δ^2 (standard error)
5%	EM	5.895E-02 (4.489E-02)	1.165E-02 (8.898E-03)	9.344E-04 (1.128E-03)
	EMB	5.889E-02 (4.353E-02)	1.163E-02 (8.614E-03)	8.523E-04 (1.055E-03)
10%	EM	2.283E-02 (1.721E-02)	4.389E-03 (3.285E-03)	5.323E-04 (4.367E-04)
	EMB	2.240E-02 (1.690E-02)	4.366E-03 (3.265E-03)	4.339E-04 (3.334E-04)
20%	EM	2.365E-02 (1.771E-02)	4.391E-03 (3.303E-03)	8.042E-04 (6.577E-04)
	EMB	2.258E-02 (1.712E-02)	4.299E-03 (3.244E-03)	6.466E-04 (4.948E-04)
30%	EM	2.432E-02 (1.880E-02)	4.377E-03 (3.344E-03)	1.064E-03 (8.295E-04)
	EMB	2.337E-02 (1.738E-02)	4.358E-03 (3.255E-03)	8.284E-04 (6.307E-04)

In summary, results of the simulation studies suggest that imputing missing values using both the EM and the EMB algorithm are good, with the EMB algorithm outperforms the EM algorithm for models of the linear functional relationship type as they give smaller values of MAE, RMSE, and smaller values of the standard error of the estimated bias in the parameters.

The EM algorithm has largely been used in solving maximum-likelihood parameter estimation problems (Dempster et al., 1977; Bilmes, 1998; Bock & Murray, 1981). It has also become popular in handling missing data because of its simplicity, in spite of its slow convergence rate (Couvreur, 1996). Nevertheless, EM has wide application in addressing missing data in medical data (Dziura et al., 2013) and environmental data (Razak et al., 2014; Zainuri et al., 2015).

In this paper, the EM algorithm is improved by integrating bootstrap in the EM procedure. Simulation studies indicate the superiority of the EMB algorithm in both LFRM1 and LFRM2 models. The re-sampling method of EMB made the estimator improved by creating a multiply-imputed values for each missing data. As a result, the average value of the imputed data set contributes towards making the estimates more accurate with smaller standard errors.

6.6 Application to Real Data

To illustrate with a practical example, a data set which consists of 96 observations that are free from any outliers (Goran et al. 1996) is considered. The study was to examine the accuracy of some widely used body-composition techniques for children, using the dual-energy X-ray absorptiometry (DXA). The sample comprises of children ages from four to ten years. They assessed the children's body fat by using two variables, namely the skinfold thickness (ST) and bioelectrical resistance (BR). The assumption made is

that the measurement error can take place in either variable of this experiment and the relationship between these two variables can be expressed in a LFRM as given in (2.1).

In the interest of measuring the performance of the EM algorithm and the EMB algorithm, the dependent variable is randomly made missing at 5%, 10%, 20%, and 30% respectively. Both LFRM1 and LFRM2 models are applied in this experiment. Table 6.9 shows the values of MAE and RMSE for LFRM1, using both imputation methods of EM and EMB. It can be seen that there is a consistency in the results whereby the EMB algorithm has smaller MAE and RMSE values as compared to using the EM algorithm. Similar conclusion can be made for the results in Table 10, in which the values of bias using the EMB algorithm are smaller in comparison to the EM algorithm.

Table 6.9: MAE and RMSE for LFRM1 for real data using two imputation methods

Percentage of missing (%)	Performance Indicator / Method	MAE	Percentage change of MAE (%)	RMSE	Percentage change of RMSE (%)
5%	EM	5.2256	10.97	4.6518	28.20
	EMB	4.6521		3.3400	
10%	EM	5.5593	27.81	5.3013	6.14
	EMB	4.0135		4.9756	
20%	EM	4.9928	26.13	4.9781	5.25
	EMB	3.6883		4.7166	
30%	EM	5.1355	20.27	5.5337	5.95
	EMB	4.0946		5.2044	

Table 6.10: Estimated bias of parameters using LFRM1 for real data

Percentage of missing (%)	Parameters / Methods	α	β	σ_{δ}^{2}
5%	EM	0.4926	0.0997	0.0782
	EMB	0.3975	0.0997	0.0573
10%	EM	0.6098	0.0997	0.1821
	EMB	0.4895	0.0997	0.1036
20%	EM	0.5243	0.0997	0.1625
	EMB	0.4366	0.0997	0.0772
30%	EM	0.6315	0.0997	0.1017
	EMB	0.6236	0.0997	0.0524

As mentioned earlier in Section 6.3, the LFRM model is considered where the slope parameter is estimated using a nonparametric method, namely LFRM2. Table 6.11 indicates the MAE and RMSE values of the slope for LFRM2 while Table 6.12 illustrates the EB of the parameters of the slope for LFRM2. From both tables, it is noted that the EMB algorithm proves to be better with smaller values of EB, MAE and RMSE.

Table 6.11: MAE and RMSE for LFRM2 for real data using two imputation methods

Percentage of missing (%)	Performance Indicator / Method	MAE	Percentage change of MAE (%)	RMSE	Percentage change of RMSE (%)
5%	EM	3.6671	39.49	5.5610	40.53
	EMB	2.2190		3.3070	
10%	EM	2.7472	12.79	3.7241	4.69
	EMB	2.3959		3.5494	
20%	EM	2.6680	36.36	5.2740	15.77
	EMB	1.6978		4.4424	
30%	EM	3.2698	18.21	3.8403	25.61
	EMB	2.6744		2.8568	

Table 6.12: Estimated bias of parameters for LFRM2 for real data

Percentage of missing (%)	Parameters / Methods	α	β	σ_δ^2
5%	EM	0.0236	0.0080	0.0963
	EMB	0.0067	0.0080	0.0128
10%	EM	0.0508	0.0080	0.1865
	EMB	0.0406	0.0080	0.0196
20%	EM	0.1538	0.0080	0.1775
	EMB	0.0373	0.0080	0.1194
30%	EM	0.1950	0.0080	0.2707
	EMB	0.1543	0.0080	0.1381

It can be inferred that from this practical application, both methods of imputations namely the EM and the EMB algorithm demonstrate good results based on the measurement of EB, MAE and RMSE values. It is shown that imputing missing values using EMB gives a better approach than the EM in handling missing values for data that can be modelled by the linear functional relationship formulation. In this practical example, it is proven that EMB has improved the precision in the algorithm and this is reflected by its superior performance.

6.7 Summary

In this chapter, two modern approaches of handling missing values have been investigated, namely the EM and the EMB algorithm for datasets that can be modelled by the linear functional relationship model. Results from the simulation study suggest both methods of imputation can be applied for two forms of the linear functional relationship model. Even in the presence of high percentage of missing values (to as high as 30%), both methods adequately handle the problem. These can be seen with small bias measure of parameter and small MAE and RMSE. When comparing the two imputation methods, EMB is superior to the EM. Again, this is evidenced by the MAE and RMSE values. EMB has several advantages where it can be easily applied to LFRM, the bootstrapping method gives better precision to the parameter estimates, and the computational time is practically fast.

A real data set that compares the relationship between two variables measurements have been illustrated. The results obtained shows that if in the case when the real data set has missing values for a percentage to as high as 30%, both methods of imputation are suitable for handling missing values with EMB being superior than EM.

CHAPTER 7: CONCLUSION AND FURTHER WORKS

7.1 Conclusion and summary

The primary goal of this research was to study the parameter estimation, outlier detection and missing values imputation in the LFRM. All four objectives of this study have been successfully achieved. For the first objective on the parameter estimation as given in Chapter 3, a robust technique using the nonparametric estimation approach is proposed to estimate the slope parameter in a LFRM. Results from the simulation study showed that the proposed method outperforms the MLE and the nonparametric method as proposed by Al-Nasser and Ebrahem (2005), by having smaller mean square error of the slope parameter and smaller estimated bias of the parameters. As for the application in real data, the proposed method suggests a more accurate estimation by having smaller values of standard deviation as compared to the MLE and the method proposed by Al-Nasser and Ebrahem (2005).

For the second objective, the $COVRATIO$ statistic is proposed to identify a single outlier in LFRM as given in Chapter 4. The cut-off points of the $COVRATIO$ statistic is determined to obtain the 1%, 5% and 10% upper percentiles respectively by using the Monte Carlo simulation method. The observation that exceeds this cut-off points are identified as outliers. From the simulation study and application to real data, the cut-off point at 5% level of significance is obtained by $y = 135.63n^{-1.145}$ and this cut-off point successfully identifies the presence of a single outlier for the data that can be modelled by LFRM.

In the third objective as explained in Chapter 5, an efficient procedure to identify the multiple outliers in a LFRM is proposed. The single linkage algorithm with the Euclidean distance as the similarity measure is used and a new stopping rule to identify the potential outliers is proposed. Here, a robust stopping rule is proposed by using the

median and median absolute deviation (MAD) of the tree heights. Results from the simulation study suggest that this proposed method successfully identifies the planted outliers in different sample sizes with different mean of contamination of the outliers. At a higher level of contamination, this proposed clustering method for the LFRM gives a high value of "success" probability (*pop*) and a low value of masking error (*pmask*) and swamping error (*pswamp*) and this indicates that this method is efficient in detecting multiple outliers. When applying to real data sets, it also proves that the proposed new stopping rule successfully detects the outliers as found in other classical data sets.

Finally, for the last objective on the missing value problem in LFRM as given in Chapter 6, two modern imputation techniques are proposed, namely the expectation-maximization (EM) algorithm and the expectation-maximization with bootstrapping (EMB) algorithm. The simulation study and real data application reveal that both of these methods are feasible in handling the missing value problem in LFRM with the EMB method being superior to the EM method.

7.2 Contributions

Several contributions from this study are given here. First, the proposed robust nonparametric estimation to estimate the slope parameter in a LFRM is a new method and it is robust to outliers. This nonparametric estimation approach makes it appealing as it does not require any assumption on the probability distribution of the data and this method is easy to apply.

Next, the identification of outliers in a LFRM is a new topic and has not been explored. With the cut-off point obtained using the $COVRATIO$ statistic, researchers will be able to identify a single outlier in a much easier way. As for the multiple outliers identification, a new stopping rule is developed using the median and median absolute deviation (MAD) of the tree heights, and this stopping rule is efficient as it is able to

identify outliers and having low values of masking error probability (*pmask*) and swamping error probability (*pswamp*). This stopping rule is novel and has an advantage as it is robust to outliers.

Finally, for the missing value problem in LFRM, the EMB algorithm is superior to the EM. This bootstrapping method gives better precision to the parameter estimates and as the computational time is practically short, it makes the EMB algorithm more appealing and is a good alternative to handle missing values for data that can be modelled by the linear functional relationship model.

7.3 Limitation of the Study and Further Works

In Chapter 3 of this research, the focus was only on estimating the slope parameter, β of the LFRM using the nonparametric approach. Further research can be done on estimating the other parameters of the LFRM, such as the error of the variance for this model which are, σ_{δ}^2 and σ_{ε}^2 as not much study has been done on the following parameters. As the slope parameter estimation in this chapter is based on an unreplicated LFRM, where there is only a single observation for each level of i as in (2.5), further investigation may be valuable to study for the replicated form and the simultaneous form of the LFRM, as mentioned by Fuller (1987). This work that is based on the simple LFRM can also be extended to the multiple LFRM.

In the identification of outliers in a LFRM, from Chapter 4 and Chapter 5, this study only examined situations when the multiple outlier is at the response variable, y. Further works can be extended by taking into account other outlier scenarios, such as outliers in the x variables or outliers in both x and y variables. The distance measure used in this clustering algorithm is the Euclidean measure. It would be a possible area of research to compare the performance of this Euclidean distance with other measurement

distances, such as the Manhattan distance and the Malahanobis distance in identifying outliers in the LFRM. Moreover, this study begins by considering agglomerative hierarchical clustering algorithm to cluster the observations. Future research might explore other advanced techniques in identifying multiple outliers such as the K-means cluster technique and by considering different stopping rules.

As for the missing value techniques in Chapter 6, the data are only considered when missing at random (MAR). It will be worthwhile to discuss further with other different missing mechanism, such as when the data are missing completely at random (MCAR), and when the data are not missing at random (MNAR). It is shown in this chapter that the EMB algorithm is superior to the EM algorithm. Another possible future work would be comparing the performance of this EMB algortithm with other modern imputation techniques, such as the Markov Chain Monte Carlo and Fully Conditional Specification (Baraldi & Enders, 2010).

REFERENCES

Abuzaid, A., Mohamed, I., Hussin, A.G., & Rambli, A. (2011). *COVRATIO* statistics for simple circular regression model. *Chiang Mai J.Sci.*, *38*(3), 321-330.

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67*: 1012–1028.

Adcock, R. J. (1877). Note on the method of least squares. *The Analyst, 4*(6), 183-184.

Adcock, R. J. (1878). A problem in least squares. *The Analyst, 5*(2), 53-54.

Adnan, R., Mohamad, M. N., & Setan, H. (2003). Multiple outliers detection procedures in linear regression. *Matematika, 19*, 29–45.

Agulló, J. (2001). New algorithms for computing the least trimmed squares regression estimator. *Computational Statistics & Data Analysis, 36*(4), 425-439.

Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster analysis. Sage University paper series on quantitative applications in the social sciences, 7-44.

Allison, P. D. (2003). Missing data techniques for structural equation modelling. *Journal of abnormal psychology, 112*(4), 545.

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. *Pakistan Journal of Statistics, 21*(3), 265-274.

Anderson, T.W. (1976). Estimation of linear functional relationship: approximate distributions and connections with simultaneous equations in econometrics (with discussion). *J.R. Statist. Soc. B, 38*, 1-36.

Baraldi, A. N. & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48,* 5-37.

Barnet, V. & Lewis, T. (1994). *Outliers in statistical data*. John Wiley, Chichester.

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics, 5*(3), 207-212.

Barzi, F. & Woodward, M. (2004). Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology, 160*(1): 34-45.

Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley, New York.

Bilmes, J.A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute,* 2-7.

Bock R.D. & Murray A. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika 46*(4): 443-459.

Buonaccorsi, J. P. (1996). A modified estimating equation approach to correcting for measurement error in regression. *Biometrika, 83*(2), 433-440.

Cai, T. T., & Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics, 34*(5), 2159-2179.

Carroll, R. J. (2005). Measurement error in epidemiologic studies. *Encyclopedia of biostatistics*. doi: 10.1002/0470011815.b2a03082.

Cateni, S., Colla, V., & Vannucci, M. (2008). Outlier detection methods for industrial applications. *Advances in Robotics, Automation and Control*. Jesus Aramburo and Antonio Ramirez Trevino (Ed.).

Chen, Z., Fu, A. & Tang, J. (2002). Detection of outliered patterns. Dept. of CSE, Chinese University of Hong Kong.

Cheng, C. L. & Van Ness, J.W. (1994). On estimating linear relationships when both variables are subject to errors. *J. R. Statist. Soc. B, 56*, 167-183.

Cheng, C., & Van Ness, J. (1999). Statistical Regression with Measurement Error. Arnold: London.

Chowdhury, G. (2010). *Introduction to modern information retrieval*: Facet publishing.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association, 74*(365), 169-174.

Copas, J. (1972). The likelihood surface in the linear functional relationship problem. *Journal of the Royal Statistical Society. Series B (Methodological)*, 274-278.

Couvreur C. (1997). The EM algorithm: A guided tour. *Computer Intensive Methods in Control and Signal Processing*, 209-222: Springer.

Dasgupta, S., & Long, P. M. (2005). Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences, 70*(4), 555-569.

de Souto, M. C., Jaskowiak, P. A., & Costa, I. G. (2015). Impact of missing data imputation methods on gene expression clustering and classification. *BMC Bioinformatics, 16*(1), 1-9.

Deming, W. E. (1931). The application of least squares. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 11*(68), 146-158.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.

Dent, B. M. (1935). On observations of points connected by a linear relation. *Proceedings of the Physical Society, 47*(1), 92.

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.

Dorff, M., & Gurland, J. (1961). Estimation of the parameters of a linear functional relation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 160-170.

Drion, E. (1951). *Estimation of the parameters of a straight line and of the variances of the variables, if they are both subject to error.* Paper presented at the Indagationes Mathematicae (Proceedings).

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

Dziura, J.D., Post, L.A., Zhao, Q., Fu, Z. & Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: From design to analysis. *The Yale Journal of Biology and Medicine 86*(3):343-358.

Elfessi, A. & Hoar, R.H. (2001). Simulation study of a linear relationship model between two variables affected by errors. *Journal of Statatistical Computation and Simulation*, *71*, 29-40.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, *4*(1), 65-75.

Everitt, B. S. (1993). *Cluster Analysis*. London: Wiley.

Faria, S., & Soromenho, G. (2010). Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, *80*(2), 201-225.

Fernholz, L. T., Morgenthaler, S., & Tukey, J. W. (2004). An outlier nomination method based on the multihalver. *Journal of Statistical Planning and Inference*, *122*(1), 125-139.

Fuller, W.A. (1987). *Measurement error models.* Wiley and Sons, New York.

Geary, R. C. (1949). Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown. *Econometrica: Journal of the Econometric Society*, 30-58.

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

George, N.I., Bowyer, J.F., Crabtree, N.M., & Chang, C.W. (2015). An iterative leave-one-out approach to outlier detection in RNA-seq data. *PLoS ONE*, *10*(6).

Gillard, J.W. (2007). *Errors in variables regression: What is the appropriate model?* (doctoral dissertation). Retrieved from http://ethos.bl.uk.

Gillard, J.W. & Iles, T.C. (2005). *Method of moment's estimation in linear regression with errors in both variables.* (doctoral dissertation). Retrieved from http://ethos.bl.uk.

Gold, M.S. & Bentler, P.M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and Expectation-Maximization. structural equation modelling: *A Multidisciplinary Journal, 7*(3), 319-355.

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.R., & Hunter, G.R. (1996). Cross-calibration of body-composition techniques against dual-energy X-Ray absorptiometry in young children. *American Journal of Clinical Nutrition, 63*, 299- 305.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. *Handbook of psychology, 1*(4), 87–114.

Guan, N.C. & Yusoff, N.S.B. (2011). Missing values in data analaysis: Ignore or impute? *Education in Medicine Journal*, *3*(1), 6-11.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, 54*(3), 761-771.

Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association, 88*(424), 1264-1272.

Hadi, A. S., Imon, A. H. M. R. & Werner, M. (2009). Identification of outliers, *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*, 57 – 70.

Hajek, J. (1969). *A course in nonparametric statistics* (Vol. 1969): Holden-Day San Francisco.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

Hawkins, D. M., & Olive, D. (1999). Applications and algorithms for least trimmed sum of absolute deviations regression. *Computational Statistics & Data Analysis, 32*(2), 119-134.

He, Z., Xu, X., & Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters, 24*(9), 1641-1650.

Honaker, J., King, G. & Blackwell, M. (2013). *Amelia II: A Program for missing data*. Retrieved from: http://gking.harvard.edu/amelia.

Housner, G. W., & Brennan, J. (1948). The estimation of linear trends. *The Annals of Mathematical Statistics, 19*(3), 380-388.

Howell, D. C. (2008). The analysis of missing data. *Handbook of social science methodology*, 208-224.

Huwang, L. & Yang, J. (2000). Trimmed estimation in the measurement error model when the covariate has replicated observations. *Proc. Natl. Sci. Counc*. ROC(A). *24*(5), 405-412.

Humphreys, R. M., Jones, T. J., & Gehrz, R. D. (1987). The enigmatic object variable A in M33. *The Astronomical Journal, 94*, 315-323.

Hussin, A., & Abuzaid, A. (2012). Detection of outliers in functional relationship model for circular variables via complex form. *Pakistan Journal of Statistics, 28*(2), 205-216.

Hussin, A. G., Zaid, A. H. A., Ibrahim, A. I. N., & Rambli, A. (2013). Detection of outliers in the complex linear regression model. *Sains Malaysiana*, *42*(6), 869-874.

Ibrahim, S., Rambli, A., Hussin, A.G. & Mohamed, I. (2013). Outlier detection in a circular regression model using COVRATIO statistic, *Communication in Statistics - Simulation and Computation*, *42* (10), 2272- 2280.

Imon, A.H.M.R. & Hadi, A.S. (2008). Identification of multiple outliers in logistic regression. *Commun. Stat. Theor. Methods*, *37*(11), 1697-1709.

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. *Developmental psychology*, *45*(4), 1195.

Junger, W.L. & de Leon AP. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment, 102*, 96-104.

Junninen, H., Niska, H., Tuppurrainen, K., Ruuskanen, J. & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atoms Environ*, *38*, 2895-2907.

Kendall, M., & Stuart, A. (1961). Estimation: least squares and other methods. *The Advanced Theory of Statistics, 2*, 75-97.

Kendall, M.G. & Stuart, A. (1973). *The Advance Theory of Statistics,* (Vol. 2). London: Griffin.

Kianifard, F., & Swallow, W. H. (1996). A review of the development and application of recursive residuals in linear models. *Journal of the American Statistical Association, 91*(433), 391-400.

Kim, M.G. (2000). Outliers and influential observations in the structural errors-in-variables model. *Journal of Applied Statistics, 4*, 451-460.

Koul, H. L., & Song, W. (2008). Regression model checking with Berkson measurement errors. *Journal of Statistical Planning and Inference, 138*(6), 1615-1628.

Kummell, C. H. (1879). Reduction of observation equations which contain more than one observed quantity. *The Analyst*, 97-105.

Lindley, D. V. (1947). Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society, Suppl., 9*, 218-244.

Lindley, D., & El-Sayyad, G. (1968). The Bayesian estimation of a linear functional relationships. *Journal of the Royal Statistical Society. Series B (Methodological)*, 190-202.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*: John Wiley & Sons.

Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research, 18*(2-3), 292-326.

Little, R. J. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association, 87*(420), 1227-1237.

Loureiro, A., Torgo, L., & Soares, C. (2004). *Outlier detection using clustering methods: a data cleaning application.* Paper presented at the Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association, 54*, 173-205.

Mamun, A., Zubairi, Y., Hussin, A., & Rana, S. (2016). A comparison of missing data handling methods in linear structural relationship model: evidence from BDHS2007 data. *Electronic Journal of Applied Statistical Analysis, 9*(1), 122-133.

Maronna, R. A., Martin, R. D. & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods.* England: John Wiley and Sons Ltd.

Midi, H. (2010). Robust logistic diagnostic for the identification of high leverage points in logistic regression model. *Journal of Applied Sciences, 10*(23), 3042-3050.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179.

Mirkin, B. (1998). Mathematical classification and clustering: From how to what and why *Classification, data analysis, and data highways* (pp. 172-181): Springer.

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*, 61-64.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal, 20*(4), 359-363.

Neyman, J., & Scott, E. L. (1951). On certain methods of estimating the linear structural relation. *The Annals of Mathematical Statistics*, 352-361.

Nurunnabi, A., Imon, A. R., & 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.

Pal, M. (1980). Consistent moment estimators of regression coefficients in the presence of errors in variables. *Journal of Econometrics, 14*(3), 349-364.

Patefield, W. (1977). On the information matrix in the linear functional relationship problem. *Applied Statistics*, 69-70.

Patefield, W. M. (1985). Information from the maximized likelihood function. *Biometrika, 72*, 664-668.

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.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research, 74*(4), 525-556.

Razak, N., Zubairi, Y., & Yunus, R. (2014). Imputing missing values in modelling the PM10 concentrations. *Sains Malaysiana, 43*(10), 1599-1607.

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

Rancoita, P.M.V., Zaffalon, M., Zucca, E., Bertoni, F. & Campos, C.P. (2015). Bayesian network data imputation with application to survival tree analysis. *Computational Statistics and Data Analysis*, 373- 387.

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

Rousseeuw, P. J., & Kaufman, L. (1990). *Finding Groups in Data*: Wiley Online Library.

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

Satari, S. Z. (2014). Parameter estimation and outlier detection for some types of circular model. (Unpublished doctoral dissertation). University of Malaya, Kuala Lumpur.

Satman, M. H. (2013). A new algorithm for detecting outliers in linear regression. *International Journal of Statistics and Probability, 2*(3), 101.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*: CRC press.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling psychology, 57*(1), 1-10.

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.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification*.

Sprent, P., & Smeeton, N. C. (2016). *Applied nonparametric statistical methods*: CRC Press.

Sprent, P. (1969). *Models in regression and related topics*. Methuen, London.

Takahashi, M. & Ito, T. (2013). Multiple imputation of Missing Values in Economic Surveys: Comparison of Competing Algorithms, *Proceedings 59th ISI World Statistics Congress, Hong Kong.*

Theil, H. (1992). A rank-invariant method of linear and polynomial regression analysis *Henri Theil's Contributions to Economics and Econometrics* (pp. 345-381): Springer.

Tsai, Jia-Ren (2010). Generalized confidence interval for the slope in linear measurement error model. *Journal of Statistical Computation and Simulation, 80*(8), 927-936.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management, 24*(1), 53-62.

Vansina, F., & De Greve, J. (1982). Close binary systems before and after mass transfer. *Astrophysics and Space Science, 87*(1-2), 377-401.

Van Huffel, S., & Vandewalle, J. (1991). *The total least squares problem: computational aspects and analysis* (Vol. 9): SIAM, Philadelphia.

Van Montfort, K. (1989). *Estimating in structural models with non-normal distributed variables: some alternative approaches*, DSWO Press, Leiden.

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, J. & Miao, Y. 2009. Note on the EM Algorithm in Linear Regression Model. *International Mathematical Forum, 38*, 1883-1889.

Wang, L., Zhang, Y., & Feng, J. (2005). On the Euclidean distance of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27*(8), 1334-1339.

Wong, M. Y. (1989). Likelihood estimation of a simple linear regression model when both variables have error. *Biometrika, 76*, 141-148.

Zainuri, N., Jemain, A., & Muda, N. (2015). A comparison of various imputation methods for missing values in air quality data. *Sains Malaysiana, 44*(3), 449-456.

LIST OF PUBLICATIONS AND PAPER PRESENTED

1. **Article Publication**: Ghapor, A.A., Zubairi, Y. Z., Mamun, A.S.M.A., and Imon, A.H.M.R. (2014). On detecting outlier in simple linear functional relationship model using *COVRATIO* statistic. *Pakistan Journal of Statistics*, *30*(1): 129-142.

2. **Article Publication**: Ghapor, A.A., Zubairi, Y. Z., Mamun, A.S.M.A., and Imon, A.H.M.R. (2015). A new nonparametric estimation for slope of linear functional relationship model. *Pakistan Journal of Statistics*, *31*(3): 339-350.

3. **Article Publication**: Ghapor, A.A., Zubairi, Y. Z., and Imon, A.H.M.R. (2015). Missing Value Estimation Methods for Data in Linear Functional Relationship Model. Manuscript accepted for publication by Sains Malaysiana on 9th June, 2016.

4. **Oral Presentation**: Detecting Outlier by Using *COVRATIO* Statistic in Linear Functional Relationship Model; presented at The 3rd International Conference on Mathematical Sciences held at the Putra World Trade Center, Kuala Lumpur (18th December 2013).

5. **Oral Presentation**: A Robust Method to Estimate the Slope Parameter in Linear Functional Relationship Model; presented at the 17th International Conference on Mathematical and Computational Methods in Science and Engineering, held at Kuala Lumpur (23-25 April 2015).

6. **Oral Presentation**: Missing Value Estimation Methods for Data in Linear Functional Relationship Model; presented in The 4th International Conference on Computer Science and Computational Mathematics, held at Langkawi, Malaysia (7-8 May 2015).