SIMULTANEOUS FUNCTIONAL RELATIONSHIP MODEL AND OUTLIER DETECTION USING CLUSTERING TECHNIQUE FOR CIRCULAR VARIABLES

NURKHAIRANY AMYRA BINTI MOKHTAR

MASTER OF SCIENCE

UNIVERSITI PERTAHANAN NASIONAL MALAYSIA

2016

SIMULTANEOUS FUNCTIONAL RELATIONSHIP MODEL AND OUTLIER DETECTION USING CLUSTERING TECHNIQUE FOR CIRCULAR VARIABLES

NURKHAIRANY AMYRA BINTI MOKHTAR

Thesis submitted to the Centre for Graduate Studies, Universiti Pertahanan Nasional Malaysia, in Fulfillment of the Requirements for the Degree of Master of Science (Statistics)

FEBRUARY 2016

ABSTRACT

This study focuses on simultaneous linear functional relationship model and outlier detection for circular variables in a simple linear functional relationship model. A new simultaneous model is extended from a simple linear functional relationship model for circular data proposed by Caires and Wyatt (2003) by assuming equal error variances. The maximum likelihood estimator of the parameters in the simultaneous model are obtained and the covariance between the parameters is derived using the Fisher Information Matrix. Results from the simulation study indicate that the estimated parameters have small bias. The second part of the study, an estimation of the concentration parameter for simultaneous linear functional relationship model for circular variables when the variances of the error term are assumed not to be equal. The modified Bessel function was expanded by using the asymptotic power series which in turn becomes a cubic equation of the concentration parameter. Simulation study was done the result shows that the estimated concentration parameter has smaller bias for large concentration parameter and large sample size based on performance measure of estimated bias, estimated standard error and a few other measures. The final part of the study considers the problem in detecting multiple outliers in circular variables for functional relationship model. A clustering-based procedure is developed for the predicted and residual values obtained for the Caires and Wyatt model. Single linkage of hierarchical clustering method is used to obtain a tree diagram in detecting outliers. Based on simulation study, it can be concluded that the probability of success is good with increasing n and level of contamination. The level of masking and swamping

decreases as n and κ gets bigger. In all of the proposed methods, we applied to a real data set to illustrate the applicability. The significant contribution of the study is the development of a simultaneous functional relationship model for circular variables that can be applied for both equal and unequal variance of error term. Another contribution of this study is a method of identifying multiple outliers in circular variables for functional relationship model based on the dendrogram plot.

ABSTRAK

Kajian ini memberi tumpuan kepada model hubungan fungsian linear serentak dan pengesanan titik terpencil untuk pembolehubah membulat dalam model hubungan fungsian linear mudah. Model serentak yang baru dilanjutkan daripada model hubungan fungsian linear mudah untuk data membulat yang dicadangkan oleh Caires dan Wyatt (2003) dengan menganggap varians ralat adalah sama. Penganggar kebolehjadian maksimum bagi parameter dalam model serentak diperoleh dan kovarians antara parameter diolah menggunakan Fisher Maklumat Matrix. Hasil daripada kajian simulasi menunjukkan bahawa parameter yang dianggarkan mempunyai *bias* yang kecil. Bahagian kedua kajian ini mencadangkan anggaran parameter kepekatan untuk model hubungan fungsian linear serentak untuk pembolehubah membulat dengan anggapan bahawa varians bagi ralat adalah tidak sama. Fungsi Bessel dikembangkan dengan menggunakan siri kuasa asimptot yang seterusnya menghasilkan persamaan kubik bagi parameter kepekatan. Kajian simulasi telah dilakukan dan keputusan menunjukkan bahawa penganggar parameter kepekatan mempunyai bias yang lebih kecil untuk kepekatan dan saiz sampel yang besar berdasarkan penganggar bias, penganggar ralat piawai (ESE) dan beberapa ukuran lain. Bahagian terakhir kajian ini melihat kepada masalah dalam mengesan pelbagai titik terpencil dalam pembolehubah membulat untuk model hubungan fungsian linear mudah. Prosedur yang berasaskan kelompok ini dibentuk untuk nilai-nilai yang diramalkan dan sisa yang diperolehi untuk Caires dan Wyatt model. Pautan tunggal kaedah pengelompokan hierarki digunakan untuk mendapatkan gambar rajah pokok dalam mengesan titik terpencil. Berdasarkan kajian simulasi, dapat disimpulkan

bahawa kebarangkalian kejayaan yang baik dengan peningkatan n dan tahap pencemaran. Tahap pelekat dan melanda berkurangan apabila n dan κ semakin besar. Kesemua kaedah ini kemudiannya digunakan untuk set data sebenar untuk menunjukkan kesesuaiannya. Sumbangan penting kajian ini ialah pembangunan model hubungan fungsian linear serentak untuk pembolehubah membulat yang boleh digunakan untuk kedua-dua varians yang sama dan varians tidak sama untuk ralat. Satu lagi sumbangan kajian ini adalah kaedah untuk mengenal pasti pelbagai titik terpencil berganda dalam pembolehubah bulat untuk model hubungan yang berfungsi menggunakan gambarajah dendrogram.

ACKNOWLEDGEMENT

I am grateful to Allah the most merciful. I also would like to express my sincere gratitude to my supervisors, Prof. Dr. Abdul Ghapor Hussin and Associate Prof. Dr. Yong Zulina Zubairi for their guidance, continuous encouragement and constant support in making this Master of Science (Statistics) possible. I really appreciate their guidance from the initial to the final level that enabled me to develop an understanding of this research thoroughly. Without their advice and assistance, it would be a lot tougher to completion.

Secondly, I acknowledge my sincere indebtedness and gratitude to my parents for their love, dream and sacrifice throughout my life. I am really thankful for their support, patience, and understanding that makes this work possible.

Thirdly, I would like to express very special thanks to my seniors for guiding me until this research has been successfully completed. Last but not least, I want to thank some of my friends for helping me during this study and being there always supporting me till the end of the study period.

APPROVAL

I certify that an Examination Committee has met on 15 January 2016 to conduct the final examination of Nurkhairany Amyra binti Mokhtar on her degree thesis entitled "Simultaneous Functional Relationship Model and Outlier Detection Using Clustering Technique for Circular Variables". The committee recommends that the student be awarded Master of Science (Statistics).

Members of the Examination Committee were as follows.

Muhd Zu Azhan Yahya, PhD

Professor

Faculty of Defence Science and Technology

Universiti Pertahanan Nasional Malaysia

(Chairman)

Ummul Fahri Abdul Rauf, PhD

Center for Foundation Studies

Universiti Pertahanan Nasional Malaysia

(Internal Examiner)

Azami Zaharim, PhD

Professor

Faculty of Engineering and Built Environment

University Kebangsaan Malaysia

(External Examiner)

APPROVAL

This thesis was submitted to the Senate of Universiti Pertahanan Nasional Malaysia and has been accepted as fulfilment of the requirements for the degree of Master of Science (Statistics). The members of the Supervisory Committee were as follows.

Abdul Ghapor Hussin, PhD

Professor

Faculty of Defence Science and Technology

Universiti Pertahanan Nasional Malaysia

(Main Supervisor)

Yong Zulina Zubairi, PhD

Assoc. Prof.

Mathematics Division

Center for Foundation Studies in Science

Universiti Malaya

(Co-Supervisor)

UNIVERSITI PERTAHANAN NASIONAL MALAYSIA

DECLARATION OF THESIS

Author's full name : Nurkhairany Amyra binti Mokhtar

Date of birth : 18 August 1991

Title : Simultaneous Functional Relationship Model and Outlier Detection Using Clustering Technique for Circular Variables

Academic session : 2014/2015

I declare this thesis is classified as:

- □ CONFIDENTIAL (Contains confidential information under the Official Secret Act 1972)*
- □ RESTRICTED (Contains restricted information as specified by the organization where research was done)*
- □ OPEN ACCESS I agree that my thesis to be published as onlines open access (full text)

I acknowledge that Universiti Pertahanan Nasional Malaysia reserves the right as follows.

1. The thesis is the property of Universiti Pertahanan Nasional Malaysia.
2. The library of Universiti Pertahanan Nasional Malaysia has the right to make copies for the purpose of research only.
3. The library has the right to make copies of the thesis for academic exchange.

_____	_____	_____
SIGNATURE	SIGNATURE OF SUPERVISOR	SIGNATURE OF SUPERVISOR
_____	_____	_____
I/C NUMBER	NAME OF SUPERVISOR	NAME OF SUPERVISOR
Date:	Date:	Date:

Note: *If the thesis is CONFIDENTIAL or RESTRICTED, please attach the letter from the organisation stating the period and reasons for confidentiality and restriction.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATONS

EIVM	Error-in-variables model
VM	Von-Mises distribution
ERMSE	Estimated root mean square error
ESE	Estimated standard error
n	Sample size
κ	Concentration parameter
$\bmod 2\pi$	Modulo 2π
α	Rotation parameter
μ	Mean
ε	Epsilon error term
δ	Delta error term
λ	Ratio of concentration parameters
ω	Level of contamination

CHAPTER 1

INTRODUCTION

1.1 Research Overview

Regression analysis is a popular technique used to model the relationship between an independent variable *X* and a dependent variable *Y* in an experiment. Simple linear regression model is used for a special case of relationships, namely, those that can be described by straight lines as the variables are linearly related. The general form of regression model is:

$$Y = \alpha + \beta X + \varepsilon \tag{1.1}$$

where α is the value of *y*-intercept, β is the gradient of the linear model and ε_i is the error term of y_i , given $i = 1,2, \dots, n$.

However, when the data are circular or directional, simple linear regression model is no longer appropriate (Caires and Wyatt (2003)). For example, if the data are measured in angles, they take the values on the unit of circle only. In other words, the data refers to a set of observations measured by angles in the intervals of $[0,2\pi)$ radians or $[0^\circ,360^\circ)$. One of the most obvious examples of this type of data is the wave direction.

Errors-in-Variables problems have been introduced by Adcock (1878) and then the topic has been observed continuously by other researchers. Error-in-Variables models (EIVM) are the models that account for measurement errors for both *X* and *Y* variables. Generally EIVM consist of three cases; functional EIVM where *X* variable is

assumed to be fixed, structural EIVM where *X* variable is assumed to be random and ultrastructural EIVM where the model is the synthesis between linear and structural EIVM.

This study is about functional EIVM, to be specific, on Caires and Wyatt model. There are some other types of functional relationship model studied by other researchers such as unreplicated linear functional relationship model by Hussin (1997), followed by his another model named unreplicated complex linear functional relationship model for circular variables in 1998. Later, a simultaneous circular functional relationship model was proposed by Hussin et al. (2009), extended from his previous functional relationship model in 1997. In 2014, Satari et al. proposed a new functional relationship model extended from Down and Mardia (2002) regression model (Satari et al. (2014)). Equation (1.2) describes the Caires and Wyatt linear functional relationship model for circular variables.

$$Y = \alpha + X \,(\mathrm{mod}\, 2\pi) \tag{1.2}$$

In this model, both *X* and *Y* variables are subject to random errors δ_i and ε_i, respectively. Parameter estimation and the covariance matrix are derived for this model.

Then, a new simultaneous model is extended from the Caires and Wyatt model to study the relationship between several circular variables. There are two considered cases in this study, first for equal and the second for unequal error variances. Next, a clustering technique is proposed to detect outliers in linear functional relationship

model for circular data. The power of performance of the proposed method is evaluated in the simulation studies.

1.2 Motivation of Study

Many have explored bivariate circular data to study the relationship between two variables. However, very few are studying the relationship between more than two circular variables when all of the variables are independent and subject to errors. Therefore, this study will attempt to propose a new simultaneous linear functional model to study the relationship between more than two circular variables using Caires and Wyatt model. Works on data that is based on Caires and Wyatt model has been considered for simultaneous model.

Secondly, outliers may result in gross deviation from prescribed experimental procedure in calculating the numerical value. Therefore, this study would like to discuss about a clustering technique in detecting outliers, for linear functional relationship model for circular variables.

1.3 Problem Statement

Caires and Wyatt have developed the model in 2003 to study the relationship on bivariate data of circular variables. The model is used to study the relationship between two variables of circular data. Hence, this study would like to extend the model to become a simultaneous model so that we can study the relationship between more than two variables of circular data.

Outliers are common problem in data. “Outlier” is defined as the one that appears to deviate markedly from other members of the sample in which it occurs (Grubbs (1969)). It is necessary to identify the outliers before further analysis are made. Here, we propose a simple method to detect multiple outliers for the Caires and Wyatt model using a clustering technique with single linkage of agglomerative hierarchical method. The power of performance of the methods is obtained from some simulation studies and the applicability is illustrated by a real data set.

1.4 Research Objectives

1. To propose a simultaneous linear functional relationship model between several circular variables.
2. To measure the performance of the proposed method using simulation study.
3. To propose a clustering technique to detect outliers in linear functional relationship model for circular data.

1.5 Scope of Research

The Caires and Wyatt model is extended to become a simultaneous linear functional relationship model, so that relationship between more than two circular variables can be studied using the proposed model. Again, MLE and the covariance matrix were derived, and here two cases are considered, namely for equal and unequal error variances. The parameters estimated are then evaluated by a simulation study

using SPlus statistical software. Next, single linkage of agglomerative hierarchical method is used for the detection of outliers in circular variables. Power of performance of the method is obtained by a simulation study and the method is illustrated by a real data set.

1.6 Limitation of Study

This study looks at the simultaneous linear functional relationship model. Thus, the limitation of the study is on data sets that can be modeled using linear functional relationship model, in particular, data of circular type only. Also, the distribution of data is assumed to be distributed by von Mises only.

1.7 Thesis Organisation

This thesis consists of seven main chapters. In Chapter 1, the introduction briefly explains the research overview, motivation of study, problem statement, research objectives, scope of research and thesis organisation.

In Chapter 2, literature review of general circular data, the von Mises distribution and Error in Variables Model (EIVM) are discussed. Previous works regarding circular regression models are reviewed as they are the basic ideas to functional relationship models. Also, error-in-variables models, specifically the studies of some functional relationship models by other researchers are described in this chapter. Further, the published works on outlier detection are discussed. The final part of the literature of the study is on detecting outliers and related works.