REPLICATION OF LINEAR FUNCTIONAL RELATIONSHIP MODEL FOR CIRCULAR VARIABLES

MOHD FAISAL BIN SAARI

DOCTOR OF PHILOSOPHY
(STATISTICS)

UNIVERSITI PERTAHANAN NASIONAL
MALAYSIA

2023

REPLICATION OF LINEAR FUNCTIONAL RELATIONSHIP MODEL FOR CIRCULAR VARIABLES

MOHD FAISAL BIN SAARI

Thesis submitted to the Centre for Graduate Studies, Universiti Pertahanan Nasional Malaysia, in fulfilment of the requirements for the Degree of Doctor of Philosophy (Statistics)

2023

ABSTRACT

The study involves circular data models and multiple outlier detection method. A linear functional relationship model (LFRM) with maximum likelihood parameters and covariance matrices is considered because previous studies had constraints and conditions for some parameters. This study introduces General Unreplicated LFRM (GULFRM) and General Replicated LFRM (GRLFRM). The maximum likelihood method evaluates all estimated parameters in all these models because circular data is expected to follow the von Mises distribution. Monte Carlo simulations validate the models. The simulation results show that the proposed parameter estimates produce a modest mean bias and a small squared mean error estimator, proving their effectiveness. A clustering-based pseudo-replicates model transforms unreplicated into replicated circular data. GRLFRM can analyse unreplicated data after replication. Monte Carlo simulations verify this pseudo-replicate model. The simulations show that the proposed parameter estimations have low bias, validating the model's efficacy. A method for identifying multiple outliers in circular data comes next. *COVRATIO* uses the covariance matrix determinant equation generated from GULFRM and GRLFRM to detect outlier points. The cut-off point equations use the simulation study's 5% upper percentile with 95% confidence level. Monte Carlo simulation experiments with outlier points are performed to assess method efficiency. According to simulations, as contamination level increases, the outlier points detection procedure approaches 100% detection. The study shows that *COVRATIO* is a valid strategy for detecting multiple outliers. A complete general cycle of circular data analysis is the study's key novelty. Circular data analysis has four paths. Circular data can be analysed in any form under any circumstance with this four-path cycle of circular analysis.

ABSTRAK

Kajian ini berkaitan tentang pemodelan data bulatan dan kaedah pengesanan beberapa titik terpencil. Model hubungan fungsian linear (LFRM) dengan parameter kemungkinan maksimum dan matriks kovarians digunakan kerana kajian terdahulu mempunyai kekangan dan had untuk beberapa parameter. Kajian ini memperkenalkan LFRM Umum Tidak Direplikasi (GULFRM) dan LFRM Umum Direplikasi (GRLFRM). Kaedah kemungkinan maksimum menilai semua parameter anggaran dalam model ini kerana data bulatan dijangka mengikuti taburan von Mises. Simulasi Monte Carlo mengesahkan model ini. Keputusan simulasi menunjukkan anggaran parameter yang dicadangkan menghasilkan bias yang kecil dan penganggar ralat min kuasa dua yang kecil, membuktikan keberkesanan model ini. Model replika pseudo berasaskan kluster mengubah data bulatan tidak direplikasi menjadi direplikasi. GRLFRM boleh menganalisis data yang tidak direplikasi selepas replikasi. Simulasi Monte Carlo mengesahkan model replika pseudo ini. Keputusan simulasi menunjukkan bahawa anggaran parameter yang dicadangkan mempunyai bias yang kecil, mengesahkan keberkesanan model. Seterusnya, kaedah untuk mengenal pasti beberapa titik terpencil dalam data bulat akan diperkenalkan. *COVRATIO* menggunakan persamaan penentu matriks kovarian yang dijana daripada GULFRM dan GRLFRM untuk mengesan titik terpencil. Persamaan pemotongan menggunakan peratus atas 5% kajian simulasi dengan tahap keyakinan 95%. Eksperimen simulasi Monte Carlo dengan titik terpencil dilakukan untuk menilai kecekapan kaedah. Menurut keputusan simulasi, apabila tahap pengubahsuaian meningkat, prosedur pengesanan titik terpencil menghampiri pengesanan 100%. Kajian menunjukkan bahawa *COVRATIO* ialah strategi yang sah untuk mengesan beberapa titik terpencil.

Kitaran lengkap analisis data bulatan secara umum adalah pembaharuan utama di dalam kajian ini. Analisis data bulatan mempunyai empat laluan. Data bulatan boleh dianalisis dalam sebarang bentuk dalam apa jua keadaan dengan menggunakan kitaran lengkap empat laluan analisis bulatan ini.

ACKNOWLEDGEMENTS

"In the name of Allah the All-Merciful, the All-Compassionate."

Thanks to Allah, Lord of the worlds, Protector and Ruler of Knowledge. No one deserves my warmest thanks and deepest gratitude more than Prof. Dr. Abdul Ghapor Hussin for recommending the thesis ideas and for his unwavering aid in mentoring and assisting me in every level of my study and in completing this thesis. His research experience has taught me a lot and sparks my interest in research, especially applied statistics. Thanks to my co-supervisor Dr. Mohd Syazwan Mohamad Anuar and Professor Dr. Yong Zulina Zubairi for their support.

I'm also indebted to Dr. Muhammad Hilmi Abdul Majid from Universiti Kebangsaan Malaysia (UKM), Dr. Muhammad Aslam Mohd Safari from Universiti Putra Malaysia (UPM), Dr. Mohd Iqbal Shamsudheen and Dr. Azuraini Mohd Arif from Universiti Pertahanan Nasional Malaysia (UPNM) for their aid during my lengthy years of study. All my UPNM colleagues deserve credit for creating a great research environment. On this occasion, I'd also like to thank UPNM, Universiti Malaya (UM) and the Malaysian government, notably the Ministry of Higher Education, for their financial support.

Special thanks to my beloved late father Saari Yahaya and my lovely mother Azizah Omar for all sacrifices that both of you had done. I also would like to thank my dearest siblings Nurul Azida, Nurul Asikin, Mohamad Zaini, Nurul Aini, Mohamad Fauzi, Nurul Aisyah and all my brothers and sisters in law for their spiritual support and encouragement during this study.

APPROVAL

The Examination Committee has met on **11 April 2023** to conduct the final examination of **Mohd Faisal bin Saari** on his degree thesis entitled **‘Some Replication Linear Functional Relationship Model for Circular Variables’.**

The committee recommends that the student be awarded the of **Doctor of Philosophy (Statistics)**.

Members of the Examination Committee were as follows.

Prof. Emeritus Dato' Dr. Wan Md Zin bin Wan Yunus
Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia
(Chairman)

Assoc. Prof. Dr. Ummul Fahri binti Abdul Rauf
Centre for Defence Foundation Studies
Universiti Pertahanan Nasional Malaysia
(Internal Examiner)

Prof. Ts. Dr. Mohd Rashid bin Ab Hamid
Centre for Mathematical Sciences
Universiti Malaysia Pahang
(External Examiner)

Prof. Dr. Muhammad Qaiser Shahbaz
Faculty of Sciences
King Abdul Aziz University
(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 **Doctor of Philosophy (Statistics)**. The members of the Supervisory Committee were as follows.

Prof. Dr. Abdul Ghapor Hussin
Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia
(Main Supervisor)

Dr. Mohd Syazwan Mohamad Anuar
Centre for Defence Foundation Studies
Universiti Pertahanan Nasional Malaysia
(Co-Supervisor)

Prof. Dr. Yong Zulina Zubairi
Centre for Foundation Studies in Science
Universiti Malaya
(Co-Supervisor)

UNIVERSITI PERTAHANAN NASIONAL MALAYSIA

DECLARATION OF THESIS

Student's full name	:	Mohd Faisal bin Saari
Date of birth	:	16 December 1986
Title	:	Some Replication Linear Functional Relationship Model for Circular Variables
Academic session	:	2019/2020

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

I further declare that this thesis is classified as:

☐ **CONFIDENTIAL** (Contains confidential information under the official Secret Act 1972)*

☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*

☑ **OPEN ACCESS** I agree that my thesis to be published as online 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/Dean of CGS/ Chief Librarian
861216-02-5587	Prof. Dr. Abdul Ghapor bin Hussin
IC/Passport No.	**Name of Supervisor/Dean of CGS/ Chief Librarian
Date: 19 MAY 2023	Date: 19 MAY 2023

*If the thesis is CONFIDENTAL OR RESTRICTED, please attach the letter from the organisation with period and reasons for confidentiality and restriction.
** Witness

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EIVM	-	Error-in-variable model
VM	-	Von Mises Distribution
LFRM	-	Linear Functional Relationship Model
CFRM	-	Circular Functional Relationship Model
GULFRM	-	General Unreplicated Linear Functional Relationship Model
GRLFRM	-	General Replicated Linear Functional Relationship Model
MLE	-	Maximum Likelihood Estimation
MMLE	-	Modified Maximum Likelihood Estimation
COVRATIO	-	Covariance Ratio
SLCT	-	Single Linkage Clustering Technique
EB	-	Estimated Bias
ERMSE	-	Estimated Root Mean Square Errors

LIST OF SYMBOLS

α	-	Rotation Parameter
β	-	Slope Parameter
κ_x	-	Error of Concentration Parameter on *X* Variables
κ_y	-	Error of Concentration Parameter on *Y* Variables
X_i	-	Incidental Parameter
λ	-	Ratio of Error Concentration Parameter
δ	-	Error Term on *X* Variables
ε	-	Error Term on *Y* Variables
$\log L$	-	Log Likelihood Equation
s	-	Number of Simulation
n	-	Total Sample Size
m	-	Number of Element in a Group
N	-	Number of Element in *X* Variables
M	-	Number of Element in *Y* Variables
p	-	Number of Group
$\bar{\alpha}$	-	Circular Mean
d	-	Circular Distance
$\bar{R}$	-	Mean Resultant Length
ω	-	Level of Contamination

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Research Background

Data can be categorised as linear or circular data. Linear data can be studied using the typical approach of statistical technique and is utilised on a regular basis. A set of collecting points on a circle of unit degree or radius constitutes circular data (Zubairi et al., 2008). Circular data analysis dates to the middle of the 18th century when statistical models for circular response variables were first considered. The study of circular data has helped numerous scientific domains, such as life science (Demir & Bilgin, 2019), biology (Landler et al., 2018), plant phenology (Hudson and Keatley, 2010) and environmental study (Mokhtar et al., 2021). Due to the nature of the angle, circular data cannot be formally analysed using conventional statistical methods.

The functional relationship model is a component of the error-in-variables model (EIVM), which contains fixed or deterministic underlying variables. Two other EIVM models are the structural relationship model and the ultrastructural relationship model. When the variables are random, the structural relationship model is utilised (Mamun et al., 2020). As the ultrastructural relationship model is a combination of the linear and structural relationship models, it comprises both random and fixed variables (Jinadasa and Tracy, 1990).

The differences between traditional regression and EIVM are outlined here. Standard regression assumes x value is technically observed without error and y is observed with error only for each pair of observations (x, y), but EIVM assumes both variables are observed with error. Second, unlike typical linear regression, EIVM does not separate explanatory variables from response variables. Lastly, if the objective is to predict one variable given the other, as opposed to examining the underlying relationship between the variables, regular linear regression is more suited.

When working with linear variables, it is common to utilise a linear functional relationship model (LFRM) to express the underlying relationship between the variables. The same holds true for circular or directional variables, which will be represented by the circular functional relationship model (CFRM). CFRM is more complicated than LFRM since the model includes trigonometric expressions. In order to simplify the relationship, circular variables might still be represented using a LFRM. The LFRM can be classified as unreplicated LFRM for unreplicated circular data and replicated LFRM for replicated circular data.

The LFRM for circular variables was first proposed by altering the regression model for circular variables (Hussin et al., 2006). Then, it was enhanced to unreplicated LFRM based on specific requirements and constraints, such as the ratio of error concentration parameter being fixed to one (Hassan et al., 2010), the slope parameter being fixed to one (Mokhtar et al., 2015) and the simple functional model with unequal error concentration (Mokhtar et al., 2020). The purpose of this study is to present a general unreplicated LFRM (GULFRM) for circular variables by attempting to abolish the previous conditions and restrictions. This broad approach will aid several fields of study, including business and economics (Sharif et al., 2019), the real estate industry (Chang et al., 2019), and agricultural water management (Pagliari et al., 2021).

The unidentifiability problem in an unreplicated LFRM can be avoided if the ratio of error concentration parameter is known in order to estimate the parameters (Mokhtar et al., 2017). However, those value is unknown in most actual circumstances because the information is either unavailable or not provided by field researchers. One essential strategy to avoid this difficulty is to collect this information from the sample itself (Arif et al., 2020). This is accomplished by recognizing groups from unreplicated linear data in which all parameters are identifiable and reliably approximated (Hussin et al., 2004).

The replicated LFRM can be employed by replicating observations from unreplicated data or by making replication available. The replicated LFRM can be utilized to avoid the LFRM's unidentifiability problem. Furthermore, since the error concentration parameter for X_i variable, κ_x and error concentration parameter for Y_i variable, κ_y can be estimated separately using replicated LFRM, the assumption or knowledge on the ratio of the error concentration parameter is no longer required. Since replicated LFRM is a relatively new subset of circular variables, there are few researchers working on it, yet its practical applicability may be found in fields such as meteorology (Moslim et al., 2021). This study will develop the replicated LFRM into a general replicated LFRM (GRLFRM) that may be used for circular variables.

If replication is unavailable from unreplicated data, the replicated LFRM cannot be used directly to analyse the data; nevertheless, it can be used to force group or pseudo-replicate the unreplicated data into replicated data. In other words, the concept of grouping and clustering can be extended to generate pseudo groups from the unreplicated data, which then can be used to estimate the parameters of replicated

LFRM. Hussin et al. (2004) proposed this method, but in this study the concept of pseudo-replicates will be expanded using clustering technique.

Another component of the study is the treatment of outliers in circular datasets. An outlier is a point or points of observation that deviate from the observed pattern. The distinction between outliers and non-outliers is not always obvious (Rahman et al., 2012). In meteorology, the presence of outliers in circular data may affect and diminish forecast accuracy (Rambli et al., 2016). In other words, to obtain a robust estimation of the parameters, it is essential to ensure that the dataset employed is devoid of outliers. This study discusses the multiple outlier detection approach for functionally related GULFRM and GRLFRM.

1.2 Problem Statement

In general, this study covers four LFRM-related problems. The first problem is related to enhancing the unreplicated LFRM for circular variables. Prior research on parameter estimate for unreplicated LFRM (Caires and Wyatt, 2003; Hussin, 2006; Hassan 2010; Mokhtar 2015; Mokhtar, 2020) contains limitations and restrictions. This study focuses on generalising the unreplicated LFRM by attempting to erase past constraints and restrictions to address this issue.

The second problem, estimating the parameter for unidentifiable unreplicated LFRM. Replicated LFRM can alleviate the problem of unidentifiability by replicating observations from unreplicated data or by providing replication. Prior research on replicated LFRM parameter estimation (Mokhtar et al., 2017) has shortcomings. This study generalises and improves the replicated LFRM by erasing earlier restraints and restrictions.

The third problem involves influential observations or extreme values for circular variables. Before conducting statistical analysis, it is crucial to identify outliers, as their existence in the dataset has a negative influence. Methods for detecting multiple outliers in proposed GULFRM and GRLFRM will be investigated in this study.

The final problem is categorising circular data or pseudo-replicates from unreplicated to replicated circular data. As previously stated, the unidentifiability problem in unreplicated LFRM requires an assumption to estimate the parameters. This is not identical to replicated LFRM, as it is not possible to replicate data from unreplicated data since the replication in unavailable. This study will investigate the ability of Pseudo-Replicates LFRM to force transform unreplicated circular data into replicated data.

In conclusion, this study will attempt to answer the following research questions.

1. How can the GULFRM for circular variables be developed and enhanced by overcoming historical constraints and limitations, and how effective is the Maximum Likelihood Estimation (MLE) in generating parameter estimates and decreasing bias in this model?
2. How does the MLE estimate parameters and reduce bias in the GRLFRM for circular variables? How does the GRLFRM solve the unidentifiability problem in the unreplicated model by supplying the ratio concentration parameter value?
3. When replication is unavailable, how may pseudo-replicates solve the unidentifiability problem in unreplicated LFRM? How does the suggested Single Linkage Clustering Technique (SLCT) produce pseudo-replicates for