UNIVERSITI TEKNOLOGI MARA

PARAMETER ESTIMATION AND OUTLIER DETECTION IN SIMULTANEOUS LINEAR FUNCTIONAL RELATIONSHIP MODEL FOR LINEAR VARIABLES

NUR AIN BINTI AL-HAMEEFATUL JAMALIYATUL

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Statistics)

College of Computing Informatics and Mathematics

October 2024

ABSTRACT

This study addresses the necessity of extending the bivariate linear functional relationship model (LFRM) to simultaneous LFRM to investigate relationships between multiple linear variables concurrently while considering errors in each variable. The existing bivariate LFRM only enables the study of the relationship between two linear variables at a time with error consideration. This study aims to bridge this gap by extending the parameter estimation from bivariate LFRM to simultaneous LFRM using the maximum likelihood estimation (MLE) method. This facilitates examining relationships among multiple linear variables with error consideration in each variable. The covariance matrix of the parameter estimates is derived using the Fisher information matrix. Simulation results demonstrate parameter estimation accuracy across various sample sizes, assessed through measures such as mean, estimated bias, and mean absolute percentage error (MAPE) of parameter estimates in simultaneous LFRM. Subsequently, a *COVRATIO* statistic is derived from the model's covariance matrix to detect the outliers in simultaneous LFRM. The cut-off points equation at the 1%, 5%, and 10% upper percentiles of the maximum value of the *COVRATIO* statistic is obtained through a Monte Carlo simulation study. Outliers are identified when the *COVRATIO* statistic exceeds these cut-off points. Simulation results of power performance indicate that the *COVRATIO* statistic effectively detects outliers for simultaneous LFRM. In the final segment of the study, the applicability of simultaneous LFRM is illustrated using synthetic data and Malaysia environmental datasets involving linear variables such as wind speed, humidity, and temperature. The Kolmogorov-Smirnov test and Q-Q plots support the normality of the data. The study contributes to the body of knowledge itself and offers applications for environmental analysis. Theoretical implications of this research include providing a more comprehensive approach to examining relationships between multiple linear variables while accounting for errors in each variable. This advancement enhances our understanding of multivariate relationships and error handling, which is critical in fields such as environmental science, economics, and engineering, where multiple interacting variables are common. Practical implications highlight that this enhancement improves predictive capabilities and facilitates informed decision-making across diverse domains, extending beyond outdoor activities. For instance, the ability to accurately model and predict the interplay between multiple climatic factors can lead to better resource management and policy-making in environmental analysis. Further research endeavours can explore applications of simultaneous LFRM in various fields, extending the scope of its utility.

ACKNOWLEDGEMENT

First and foremost, I express my gratitude to God for granting me the opportunity to pursue my master's degree and for guiding me through this lengthy and challenging journey to its successful completion. I extend my heartfelt thanks to my main supervisor, Dr. Nurkhairany Amyra Binti Mokhtar, for her unwavering support and guidance, which were crucial to my comprehensive understanding of this research. Without her invaluable advice and assistance, achieving this milestone would have been considerably more challenging. I would also like to acknowledge the contributions of my co-supervisors, Dr. Basri Bin Badyalina and Dr. Adzhar Bin Rambli, who provided essential support during the simulation and writing phases. My sincere appreciation goes to the MET Department Malaysia for providing the vital data necessary for this research. Special appreciation goes to my colleagues and friends for their assistance throughout this study. I dedicate this thesis to my beloved father and mother, whose vision and determination to educate me have been instrumental in my success. This achievement is a testament to both of you, and I am truly grateful for your unwavering support. Alhamdulillah, for this remarkable journey. Finally, I sincerely thank the Ministry of Higher Education Malaysia for their support through the Fundamental Research Grant Scheme (FRGS/1/2021/STG06/UITM/02/5), which significantly contributed to the success of this study.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Research Background

The error-in-variables model (EIVM) or measurement error model has been a significant subject of study for over a century, originating from early investigations into the relationships between variables (Arif et al., 2021a; Ghapor et al., 2014). The EIVM traces its roots from 1878 when Adcock aimed to find a straight line that best matched bivariate data, even though the measurements were affected by errors. Since then, extensive research has delved into various aspects of EIVM, as evident in works by (Anderson, 1976; Alshargawi et al., 2019; Arif et al., 2020a, 2022; Fuller, 1987; Ghapor et al., 2014; Mokhtar et al., 2019a). EIVM is categorised into three types: functional, structural, and ultrastructural relationship model (Mokhtar et al., 2021b). The functional relationship model applies when X is a fixed mathematical variable while the structural relationship model applies when X is a random variable. The ultrastructural relationship model emerges when both functional and structural relationships are combined (Alshargawi et al., 2019; Arif et al., 2021a; Hassan et al., 2010).

The EIVM represents an extension of the traditional regression model, encompassing scenarios where both variables X and Y are continuous and subject to measurement errors. These errors can manifest across diverse disciplines, including econometrics, environmental sciences, engineering, manufacturing, and more. For instance, in industrial settings, instrument inconsistencies might arise due to variations in the measurement process (Arif et al., 2021a). Failure to account for measurement errors significantly impacts the desired qualities of an estimator, making the EIVM more suitable than the traditional regression model in this context (Arif et al., 2019). In real life, measurement errors are prevalent in various disciplines like econometrics, environmental sciences, engineering, and manufacturing (Buonaccorsi, 1996; Doganaksoy & Van Meer, 2015).

For instance, in manufacturing, errors may result from variations in the measuring process, including operator error and instrument calibration issues (Arif et al., 2021b). Similarly, measuring household lead levels in environmental sciences involves an error-prone process due to exposure to factors like air, dust, and soil, with