

UNIVERSITI TEKNOLOGI MARA

**TYPES OF COVARIATE AND
DISTRIBUTION EFFECTS ON
PARAMETER ESTIMATES AND
GOODNESS-OF-FIT TEST USING
CLUSTERING PARTITIONING
STRATEGY FOR MULTINOMIAL
LOGISTIC REGRESSION**

HAMZAH BIN ABDUL HAMID

Thesis submitted in fulfillment
of the requirement for the degree of
Doctor of Philosophy

Faculty of Computer and Mathematical Sciences

May 2017

ABSTRACT

This thesis presents a simulation study on parameter estimation for binary and multinomial logistic regression, and the extension of the clustering partitioning strategy for goodness-of-fit test to multinomial logistic regression model. The motivation behind this study is influenced by two main factors. Firstly, parameter estimation is often sensitive to sample size and types of data. Simulation studies are useful to assess and confirm the effects of parameter estimation for binary and multinomial logistic regression under various conditions. The first phase of this study covers the effect of different types of covariate, distributions and sample size on parameter estimation for binary and multinomial logistic regression model. Data were simulated for different sample sizes, types of covariate (continuous, count, categorical) and distributions (normal or skewed for continuous variable). The simulation results show that the effect of skewed and categorical covariate reduces as sample size increases. The parameter estimates for normal distribution covariate apparently are less affected by sample size. For multinomial logistic regression model with a single covariate, a sample size of at least 300 is required to obtain unbiased estimates when the covariate is positively skewed or is a categorical covariate. A much larger sample size is required when covariates are negatively skewed. In Phase 2, we investigate the goodness-of-fit (GoF) tests for multinomial logistic regression. Goodness-of-fit tests are important to assess if the model fits the data. We investigated the Type I error and power of two goodness-of-fit tests for multinomial logistic regression via a simulation study. The GoF test using partitioning strategy (clustering) in the covariate space, χ_{p*G}^2 was compared with another test, C_g which was based on grouping of predicted probabilities. The power of both tests was investigated when quadratic term or interaction term were omitted from the model. The proposed test χ_{p*G}^2 shows good Type I error and ample power except for multinomial models with highly skewed covariate distribution. Additionally, the proposed test χ_{p*G}^2 has good power in detecting omission of continuous interaction term. Further simulation results show that partitioning strategy using Hierarchical Clustering with Canberra distance, χ_{c*G}^2 performs better than χ_{p*G}^2 (Hierarchical clustering with Euclidean distance) and χ_{k*G}^2 (Partitioning using k-medoids). The application on a real dataset confirmed the simulation results. The simulation and analyses were carried out using R, an open-source programming language for statistical computing and graphics.

ACKNOWLEDGEMENT

In the Name of Allah, the All-Compassionate, All-Merciful

Alhamdulillah, praise to Allah S.W.T for giving me the strength and opportunity to embark on my PhD and for completing this long and challenging journey successfully. My utmost gratitude and thanks go to my main supervisor Professor Dr. Yap Bee Wah for her constant guidance, support and encouragement. Her kindness, meticulousness to details, brilliant ideas and her exemplary passion for research gave me the strength and inspired me to explore this field, which I have started with zero knowledge.

I would like to thank my co-supervisors, Professor Dr. Xian-Jin Xie for his ideas, support and commitment. My special thanks also go to Assoc. Prof. Dr. Mohd Tahir Ismail and Professor Dr. Ong Seng Huat. Thank you for assisting me with this research. Thank you to Dr. Norazan Mohamed Ramli and Dr. Norshahida Shaadan for their valuable advice and comments. Thank you to all my colleagues and friends for their never-ending support, advice and helping me with this research, especially to Ainur Amira Kamaruddin, Hezlin Aryani Abdul Rahman, Nor Amirah Busul Aklan and Sahimel Azwal Sulaiman.

Finally, this thesis is dedicated to my very dear father and mother, Hj. Abdul Hamid bin Hanapi and Hj. Faridah binti Nainar. Thank you very much for providing me the best education since the beginning and remain my inspiration throughout. This piece of work is specially dedicated to both of you. Special thanks also to all my brothers and sister for their continuous support, especially Dr. Hazrul bin Abdul Hamid. *Alhamdulillah.*

I thank you all with doa, Jazaka-allahu Kayran (May Allah bless you with goodness).

TABLE OF CONTENTS

	Page
CONFIRMATION BY PANEL OF EXAMINERS	ii
AUTHOR'S DECLARATION	iii
ABSTRACT	iv
ACKNOWLEDGEMENT	v
LIST OF TABLES	x
LIST OF FIGURES	xiii
LIST OF ABBREVIATIONS	xv
CHAPTER ONE: INTRODUCTION	1
1.1 Background of Study	1
1.2 Problem Statement	5
1.3 Research Questions	7
1.4 Objectives	7
1.5 Scope of Research	8
1.6 Organization of Thesis	8
CHAPTER TWO: LITERATURE REVIEW	10
2.1 Introduction	10
2.2 General Linear Model	10
2.3 Generalized Linear Model	12
2.4 Logistic Regression Model	14
2.4.1 Binary Logistic Regression	14
2.4.2 Multinomial Logistic Regression	16
2.4.3 Ordinal Logistic Regression	18
2.5 Applications of Logistic Regression Model	19
2.5.1 Applications of Binary Logistic Regression	20
2.5.2 Applications of Multinomial Logistic Regression	21
2.5.3 Applications of Ordinal Logistic Regression	25

2.6	Issues in Parameter Estimation	28
2.7	Goodness-of-fit Tests	29
2.7.1	Goodness-of-fit Tests for Binary Logistic Regression	30
2.7.2	Goodness-of-fit Tests for Multinomial Logistic Regression	35
2.7.3	Goodness-of-fit Tests for Ordinal Logistic Regression	37
2.8	Simulation Studies	45
2.8.1	Simulation Studies on Goodness-of-fit Tests for Binary Logistic Regression	45
2.8.2	Simulation Studies on Goodness-of-fit Tests for Multinomial Logistic Regression	46
2.8.3	Simulation Studies on Goodness-of-fit Tests for Ordinal Logistic Regression	46
2.9	Distance Metrics	51
2.10	Clustering Techniques	54
2.10.1	Hierarchical Methods	55
2.10.2	Partitioning Methods	57
2.10.3	Density-Based Methods	59
2.10.4	Grid-Based Methods	60
2.10.5	Model-Based Methods	60
2.11	Applications of clustering techniques	62
2.12	Summary	63
 CHAPTER THREE: RESEARCH METHODOLOGY FOR SIMULATION STUDY		 64
3.1	Introduction	64
3.2	Phase I: Investigating the Effect of Covariates and Sample Size on Parameter Estimation	64
3.2.1	Binary Logistic Regression Model	64
3.2.2	Multinomial Logistic Regression Model	68
3.3	Phase II: Investigating The Power of Goodness-of-fit tests For Multinomial Logistic regression	71
3.3.1	Goodness-of-fit Test Based on Grouping of Predicted Probabilities Strategy	71