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^2_{p*G} 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^2_{p*G} shows good Type I error and ample power except for multinomial models with highly skewed covariate distribution. Additionally, the proposed test χ^2_{p*G} has good power in detecting omission of continuous interaction term. Further simulation results showd that partitioning strategy using Hierarchical Clustering with Canberra distance, χ^2_{c*G} performs better than χ^2_{p*G} (Hiearchical clustering with Euclidean distance) and χ^2_{k*G} (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 gratitute 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 tou 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 Hjh. 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