UNIVERSITI TEKNOLOGI MARA

THE ENHANCEMENT OF PARTIAL ROBUST M-REGRESSION (PRM) AND SPLIT SAMPLE BOOTSTRAP (SSB) FOR HIGH DIMENSIONAL DATA WITH OUTLIERS

MAZNI BINTI MOHAMAD

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Computer and Mathematical Sciences

February 2019

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Mazni Binti Mohamad
Student I.D. No.	:	2011218262
Programme	:	Doctor of Philosophy (Statistics) – CS990
Thesis Title	:	The Enhancement of Partial Robust M-Regression (PRM) and Split Sample Bootstrap (SSB) for High Dimensional Data with Outliers
Signature of Student	:	…….
Date	:	February 2019

ABSTRACT

Partial Least Squares regression (PLSR) is a regression technique that is commonly used to analyse the relationship between variables in high dimensional data. PLSR also offers good solution to multicollinearity problem in regression analysis. Hence, PLSR is a very powerful tool in dealing with multivariate data especially that of high dimension. Multivariate data is however often contains outliers. The presence of outliers in a data set may cause the regression parameter estimates become imprecise; hence, lead us to making such an invalid conclusion. Several robust PLSR methods have been proposed by researchers to cater this outlying problem. One of those is known as Partial Robust M-regression (PRM). This method is very much emphasizes on the robust starting values and the weights to be used. Reason for such emphasis is to ensure that more protection can be given against both vertical outliers and high leverage points. This study focuses on the enhancement of PRM method by introducing several PRM-based methods. Altogether there are five PRM-based methods that have been proposed in this research which are Winsorized Mean PRM-based method (PRMWM), Tukey Bisquare PRM-based method (PRMBS), Hampel PRM-based method (PRMH), the integrated Winsorized Mean and Tukey Bisquare PRM-based method (PRMWMBS) and the integrated Winsorized Mean and Hampel method (PRMWMH). The performances of all methods are assessed through both numerical examples and simulation studies under various outlying conditions. In most conditions, the PRMWM, PRMBS and PRMH outperform the original PRM. However, the integrated approach seems not to be a good idea as the two methods proposed in this study which are PRMWMBS and PRMWMH are not performing well compared to PRM. In short, this study perceives that PRMWM is the best method among all proposed methods because it consistently outperforms other methods. In order to further assess the performance of the methods, this study introduced a new bootstrap method for regression models with high dimensional data which is Weighted Split Sample Bootstrap (WSSB). The accuracy of this bootstrap technique has been measured and compared to that of the Split Sample Bootstrap (SSB). Results indicate that WSSB outperforms SSB. Therefore, WSSB is then used to compare the performance of PRMWM as compared to PRM via numerical examples and simulation studies. In general, it can be seen that PRMWM outperforms other methods as it produces the smallest prediction error values.

ACKNOWLEDGEMENT

Praise to Allah the Most Gracious and Most Merciful. With His blessings, I finally managed to complete this study. I would like to take this opportunity to express my gratitude to all those who gave me the possibility to complete this thesis.

I am deeply indebted to my supervisor, Dr. Norazan Mohamed Ramli for the continuous support of my PhD study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my PhD study. I would also like to thank my co-supervisor, Associate Professor Dr. Nor Azura Md Ghani@Mamat for her insightful comments and encouragement throughout my PhD journey as well as in my research works.

I am gratefully acknowledged the Ministry of Higher Education Malaysia and Universiti Teknologi MARA (UiTM) for the financial support given to me throughout the accomplishment of my study.

My sincere thanks also goes to all my friends and colleagues at the Faculty of Computer and Mathematical Sciences, UiTM, who give supports and help me a lot in making this thesis possible.

Finally, I would like to dedicate this work to my family - my beloved husband, Mohd Noor; my children, Firdaus, Furqan, Fatihah, Firzanah and Farisah; my late father, Hj Mohamad and my mom, Hjh Che Pah for supporting me spiritually throughout writing this thesis and my life in general. They all kept me going, and this thesis would not have been possible without them.

TABLE OF CONTENTS