UNIVERSITI PUTRA MALAYSIA

ROBUST DIAGNOSTICS IN LOGISTIC REGRESSION MODEL

SYAIBA BALQISH BINTI ARIFFIN @ MAT ZIN

FS 2010 19

ROBUST DIAGNOSTICS IN LOGISTIC REGRESSION MODEL

By

SYAIBA BALQISH BINTI ARIFFIN @ MAT ZIN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia,
in Fulfilment of the Requirements for the Degree of Master of Science

April 2010

To my noblest parents,

Haji Ariffin @ Mat Zin
Hajah Syarqiah

…who had always believed in the importance of knowledge.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia
in fulfilment of the requirement for the degree of Master of Science

ROBUST DIAGNOSTICS IN LOGISTIC REGRESSION MODEL

By

SYAIBA BALQISH BINTI ARIFFIN @ MAT ZIN

April 2010

Chairman: Habshah Midi, PhD

Faculty: Science

In recent years, due to inconsistency and sensitivity of the Maximum Likelihood Estimator (MLE) in the presence of high leverage points and residual outliers, diagnostic has become an essential part of logistic regression model. High leverage points and residual outliers have huge tendency to break the covariate pattern resulting in biased parameter estimates. The identification of high leverage points and residual outliers are believed to be vital in order to improve the performance of the MLE.

The presence of high leverage points and the residual outliers give adverse effect on the inferences by inducing large values to the Influence Function (IF). For the identification of high leverage points, Imon (2006) proposed the Distance from the Mean (DM) diagnostic method. The weakness of the DM method is that it tends to swamp some low leverage points even though it can identify the high

leverage points correctly. Deleting the low leverage points may lead to a loss of efficiency and precision of the parameter estimates.

The Robust Logistic Diagnostic (RLGD) is proposed as an alternative approach that performs well compared to the DM method. The RLGD method incorporates robust approaches and diagnostic procedures. Robust approach is firstly used to identify suspected high leverage points by computing the Robust Mahalanobis Distance (RMD) based on Minimum Volume Ellipsoid (MVE) estimator or Minimum Covariance Determinant (MCD) estimator. For confirmation, the diagnostic procedure is used to compute potential. The RLGD method ensures only correct high leverage points are identified and free from the swamping and masking effects. The performance of the RLGD method is investigated by real examples and the Monte Carlo simulation study. The real examples and the simulation results indicate that the RLGD method correctly identify the high leverage points (increase the probability of the Detection of Capability (DC)) and manage to reduce the number of swamping low leverage points (decrease the probability of the False Alarm Rate (FAR)).

The Standardized Pearson Residual (SPR) only successful in identifying a single residual outlier. The SPR method is less effective when residual outliers are present in the covariates. The Generalized Standardized Pearson Residual (GSPR) proposed by Imon and Hadi (2008) is a successful method in identifying residual outliers. However, in the initial stage of the GSPR method utilizes the graphical methods which are based on the observation’s judgement and not suitable for higher dimensional covariates. The Modified Standardized Pearson Residual (MSPR) based on the RLGD method is proposed which is more reliable. The MSPR method provides an alternative method to the GSPR method that produces similar result. The attractive feature of the MSPR method is that it is easier to apply.

This research also utilizes the RLGD method in bootstrap procedures. The Classical Bootstrap (CB) procedure by Random-x Re-sampling is not robust to the high leverage points. To accommodate this problem, the newly develop bootstrap procedures based on the RLGD method which are called the Diagnostic Logistic Before Bootstrap (DLGBB) and the Weighted Logistic Bootstrap with Probability (WLGBP) are proposed. In the DLGBB procedure, the high leverage points are excluded before applying the re-sampling process. Meanwhile in the WLGBP procedure, the high leverage points are attributed with low probabilities and consequently having low chances of being selected in the re-sampling process. Simulation results show that the DLGBB and the WLGBP procedures are more robust to the high leverage points compared to the CB procedure.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

DIAGNOSTIK TEGUH DALAM MODEL REGRESI LOGISTIK

Oleh

SYAIBA BALQISH BINTI ARIFFIN @ MAT ZIN

April 2010

Pengerusi: Habshah Midi, PhD

Fakulti: Sains

Dalam beberapa tahun kebelakangan ini, diagnostik memainkan peranan penting dalam regresi logistik berpunca daripada ketidakkonsisten dan sensitiviti Pengganggar Kebolehjadian Maksimum (MLE) dengan kehadiran titik tinggi tuasan dan titik terpencil. Titik tinggi tuasan dan titik terpencil mempunyai kecenderungan besar dalam merubah bentuk taburan kovariat menyebabkan kepincangan dalam anggaran parameter. Pengenalpastian titik tinggi tuasan dan titik terpencil dipercayai menjadi keutamaan dalam memperbaiki prestasi MLE.

Kehadiran titik tinggi tuasan dan titik terpencil memburukkan pentakbiran dengan meningkatkan Fungsi Pengaruh (IF). Dalam pengenalpastian titik tinggi tuasan, Imon (2006) mencadangkan kaedah diagnostik Jarak dari Purata (DM). Kelemahan kaedah DM adalah cenderung memperlihatkan titik rendah tuasan sebagai titik tinggi tuasan walaupun kaedah ini boleh mengenalpasti titik tinggi

tuasan dengan tepat. Membuang titik rendah tuasan menyebabkan penganggaran parameter kurang jitu dan tepat.

Kaedah Diagnostik Logistik Teguh (RLGD) dicadangkan sebagai alternatif yang menunjukkan prestasi lebih baik berbanding dengan kaedah DM. Kaedah RLGD menggabungkan aplikasi teguh dan prosedur diagnostik. Pertama, aplikasi teguh digunakan dalam mengenalpasti titik tinggi tuasan dengan mengira Jarak Teguh Mahalanobis (RMD) berdasarkan penganggar Saiz Minimum Ellipsoid (MVE) atau penganggar Penentu Kovariat Minimun (MCD). Bagi menentusahkan, prosedur diagnostik digunakan untuk mengira potensi. Kaedah RLGD memastikan hanya titik tinggi tuasan sebenar dikenalpasti dan bebas dari kesan “swamping” dan “masking”. Prestasi kaedah RLGD dikaji menggunakan data sebenar dan kajian simulasi Monte Carlo. Keputusan daripada data sebenar dan simulasi menunjukkan kaedah RLGD dapat mengenalpasti titik tinggi tuasan dengan tepat (peningkatan kepada kebarangkalian Keupayaan Pengenalpastian (DC)) dan berupaya mengurangkan bilangan titik rendah tuasan terpilih (penurunan kepada kebarangkalian Kadar Pengenalpastian Palsu (FAR)).

Penetapan Ralat Pearson (SPR) hanya cemerlang dalam pengenalpastian satu titik terpencil. Kaedah SPR menjadi tidak cekap dengan kehadiran titik terpencil berganda dalam kovariat. Penetapan Ralat Pearson Teritlak (GSPR) dicadangkan oleh Imon dan Hadi (2008) merupakan kaedah cemerlang dalam pengenalpastian titik terpencil berganda. Walaubagaimanapun, peringkat awal kaedah GSPR menggunakan kaedah grafik yang berdasarkan penilaian secara pengamatan dan tidak sesuai bagi dimensi kovariat yang lebih tinggi. Pengubahsuaian Penetapan Ralat Pearson (MSPR) berdasarkan kaedah RLGD dicadangkan dan lebih dipercayai. Kaedah MSPR sebagai alternatif kepada kaedah GSPR yang memberikan keputusan yang sama. Kaedah MSPR juga mudah diaplikasikan.

Kajian ini juga menggunapakai kaedah RLGD dalam prosedur butstrap. Prosedur Butstrap Klasik (CB) seperti Persampelan Semula –x Secara Rawak tidak teguh dengan kehadiran titik tinggi tuasan. Bagi menyelesaikan masalah ini, prosedur butstrap baru berdasarkan kaedah RLGD dikenali sebagai Diagnostik Logistik Sebelum Butstrap (DLGBB) dan Butstrap Kebarangkalian Berpemberat Logistik (WLGBP) dicadangkan. Mengikut kaedah DLGBB, titik tinggi tuasan dibuang sebelum proses persampelan semula. Manakala bagi kaedah WLGBP, titik tinggi tuasan menerima kebarangkalian yang rendah dan mempunyai peluang yang tipis untuk terpilih dalam proses persampelan semula. Hasil simulasi menunjukkan prosedur DLGBB dan WLGBP lebih teguh dengan kehadiran titik tinggi tuasan berbanding dengan prosedur CB.

ACKNOWLEDGEMENTS

My deepest gratitude and warmest thanks goes to the chairman of supervisory committee, Associate Professor Dr. Habshah Midi for encouraging me to pursue this topic, for her patience in teaching me to deal with complex problems and for her wise guidance, motivation, support and valuable advice through the entire research. Without her help, it would be impossible for me to develop such a thesis. I would also like to extend my appreciation to the supervisory committee members, Associate Professor Dr. Kassim Haron and Associate Professor Dr. Noor Akma Ibrahim for their kind supports and suggestions. I am particularly grateful to Professor Dr. A. H. M Rahmatullah Imon from Ball State University, Muncie USA, for his comments and criticism in my works and as co-author for my presented paper. I am very much impressed to Dr. Abdul Kudus and Dr. Saroje Kumar Sarkar by their ability to solve some of my problems in theoretical and programming, by their great patience in teaching me and their quick response whenever they are asked for help.

The financial supports from Fundamental Grant Research Scheme, Ministry of Higher Education of Malaysia and Graduate Research Fellowship, Universiti Putra Malaysia are grateful acknowledged. Many thanks also go to staffs, research colleagues and friends from Kolej Teknologi Darul Naim, Kota Bharu, Institute for Biostatistics and Research Methodology, Universiti Sains Malaysia, Kubang Kerian, Department of Mathematics and Institute for Mathematical

Research, Universiti Putra Malaysia, Serdang for their generous help throughout my few years in study.

My noblest father and my great mother are the reasons of my success. I am indebted to them for all the stages left and remaining of my life. To my lovely brothers and sisters, they deserve my special recognition for their love and support.

I certify that a Thesis Examination Committee has met on 27 April 2010 to conduct the final examination of Syaiba Balqish Binti Ariffin @ Mat Zin on her thesis entitled "Robust Diagnostics in Logistic Regression Model" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Master of Science.

Members of the Thesis Examination Committee were as follows:

Mahendran Shitan, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Isa Daud, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Mohd Rizam Abu Bakar, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Ibrahim Mohamad, PhD
Associate Professor
Institute of Mathematical Sciences
Universiti Malaya
Malaysia
(External Examiner)

BUJANG KIM HUAT, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 23 July 2010

This thesis submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Habshah Midi, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Kassim Haron, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Noor Akma Ibrahim, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

HASANAH MOHD GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 12 August 2010

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Putra Malaysia or other institutions.

SYAIBA BALQISH BINTI ARIFFIN @ MAT ZIN

Date: 27 April 2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BACON	Block Adaptive Computationally Efficient Outlier Nominator
BOFOLS	Best Omitted from the Ordinary Least Squares Techniques
BY	Bianco and Yohai
CB	Classical Bootstrap
CUBIF	Conditionally Unbiased Bounded Influence Function
DBB	Diagnostic-Before-Bootstrap
DC	Detection of Capability
DLGBB	Diagnostic Logistic Before Bootstrap
DM	Distance from the Mean
DRGP	Diagnostic Robust Generalized Potentials
ESR	Erythrocyte Sedimentation Rate
FAR	False Alarm Rate
GSPR	Generalized Standardized Pearson Residual
IF	Influence Function
IRLS	Iterative Re-weighted Least Squares
LMS	Least Median Squares
LTS	Least Trimmed Squares
MALLOWS	Weighted Maximum Likelihood Estimator with Mallows Type Leverage Dependent Weights
MAD	Median Absolute Deviance
MCD	Minimum Covariance Determinant
MD	Mahalanobis Distance
MLE	Maximum Likelihood Estimator

MSPR	Modified Standardized Pearson Residual
MPC	Modified Prostate Cancer
MVE	Minimum Volume Ellipsoid
MVSD	Modified Vaso-constriction in the Skin of the Digits
OLS	Ordinary Least Squares
PB	Percentile Bootstrap
PC	Prostate Cancer
RLGD	Robust Logistic Diagnostic
RMD	Robust Mahalanobis Distance
RMSE	Root Mean Square Error
SPR	Standardized Pearson Residual
VSD	Vaso-constriction in the Skin of the Digits
WBP	Weighted Bootstrap with Probability
WBY	Weighted Bianco and Yohai
WLGBP	Weighted Logistic Bootstrap with Probability
WMLE	Weighted Maximum Likelihood Estimation

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation for this Research

In recent years, the application of logistic regression model is widely use in researches. From its original acceptance in epidemiology, the model is now commonly employed in many fields including biomedical, business and finance, criminology, ecology, engineering, health policy, linguistic and wildlife biology. At the same time, statisticians continuously put efforts in research on all statistical aspects of logistic regression model. Prior to doing research on logistic regression model, it is important to understand that the objective of an analysis using this model is the same as that of any model building technique used in statistics. We would like to find the best fitting, cost-conscious and reasonable model to describe the relationship between an outcome (dependent or response) variable and a set of predictor (independent or explanatory) variables. The predictor variables are often called covariates. What distinguish logistic regression model from linear regression model is that the outcome variable in logistic regression model is binary or dichotomous $(0,1)$. For examples, doctor and pharmacist would like to determine the association between medical treatment with the survival or death of cancer patient after being discharge from hospital, to explore the relationship between age, weight, lifestyle and family medical history of patient with the presence or absence of coronary heart disease and to investigate the effect of economic crisis with the increase or decrease of fatal rate. The difference between logistic regression model and linear regression model is reflected both in the choice of parametric model and in the assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression model follow the same general principles used in linear regression model. Thus, the techniques used in linear regression model analysis will motivate our approach to logistic regression model (see Hosmer and Lemeshow, 2000).

In any regression problem, the major quantity is the mean value of the response variable, given the value of the explanatory variables. This major quantity is called the conditional mean and will be expressed as $E(Y|X)$ where Y denotes the response variable and X denotes a value of the explanatory variables. In linear regression model, we assume that this mean maybe expressed as linear equation in X, such as. $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = X\beta.$ This expression implies that it is possible for $E(Y|X)$ to take on any value as X ranges between $(-\infty, +\infty)$. For binary response, the conditional mean lies between the ranges $0 \leq E(Y|X) \leq 1$. The change in $E(Y|X)$ per unit change in X become progressively smaller as the conditional mean gets closest to 0 or 1. It resembles a plot of a cumulative distribution of random variable. Therefore, the logistic regression model can be presented by curve with S shaped for two dimension and hyper plane in the case of higher dimensions. The logistic regression model can be written as:

$$E(Y|X) = \pi(X). \tag{1.1}$$