UNIVERSITI PUTRA MALAYSIA

ROBUST DIAGNOSTIC AND ESTIMATION FOR BINARY LOGISTIC REGRESSION MODEL IN THE PRESENCE OF MULTICOLLINEARITY AND HIGH LEVERAGE POINTS

SYAIBA BALQISH BINTI ARIFFIN @ MAT ZIN

FS 2019 41

ROBUST DIAGNOSTIC AND ESTIMATION FOR BINARY LOGISTIC REGRESSION MODEL IN THE PRESENCE OF MULTICOLLINEARITY AND HIGH LEVERAGE POINTS

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 Doctor of Philosophy

December 2018

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

ROBUST DIAGNOSTIC AND ESTIMATION FOR BINARY LOGISTIC REGRESSION MODEL IN THE PRESENCE OF MULTICOLLINEARITY AND HIGH LEVERAGE POINTS

By

SYAIBA BALQISH BINTI ARIFFIN @ MAT ZIN

December 2018

Chair : Professor Habshah Midi, PhD
Faculty : Science

The binary logistic regression model popularly used in medical data analysis. In spite of its popularity, there are only a few available robust methods for this model to encounter the effects of high leverage points and multicollinearity. Failure to address model adequacy when a combination of high leverage points and multicollinearity exist in data, lead to misleading and incorrect inferences. This study is aimed to develop new robust diagnostic and estimation for logistic regression (overlap cases) and hidden logistic regression (non-overlap cases).

A new robust diagnostic called Logistic Influential Outlier Nominator (LION) is developed to identify influential outliers and the LION successfully detect the outliers in both x and y directions. Then, second robust diagnostic, namely Diagnostic Influential Observations (DIO) is developed, specifically to identify high leverage influential observations (HLIO). The DIO introduces two important stages whereby the initial stage employs the LION procedure and the confirmation stage comprises combine measures of Generalized Distance from the Mean and Generalized Standardized Pearson Residual to flag the HLIO.

Adjusted Weighted Bianco and Yohai (AWEBY) is an improvisation on the Weighted Bianco and Yohai (WBY) robust estimator. The AWEBY is proposed to increase the efficiency of WBY estimator by constructing a "smooth rejection" to replace the "hard rejection" weight function. In the AWEBY, new robust weights are formulated based on the DIO and found to properly reduce the effect of HLIO whilst protecting the good leverage points. In combined problems of HLIO and multicollinearity for overlap cases, the AWEBY estimator

is integrated for computing robust ridge parameter and formed Robust Ridge Logistic (RRL) iterative update scheme. By using the updated robust weights, the impact of the HLIO and multicollinearity will be toned down immensely.

Adjusted Weighted Maximum Estimated Likelihood (AWEMEL) in hidden logistic regression is proposed to rectify the HLIO in separation problem. New robust weights in the AWEMEL is designed based on DIO which particularly down weighs the HLIO but not the good leverage points. Finally, Robust Ridge Hidden Logistic (RRHL) is proposed to remedy both HLIO and multicollinearity for separation problem. In RRHL's iteration, the AWEMEL estimator is employed to compute robust ridge parameter which is resistance towards the bad impacts of HLIO.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

DIAGNOSIS DAN PENGANGGARAN TEGUH BAGI MODEL REGRESI LOGISTIK BINARI DENGAN KEHADIRAN MULTIKOLINEARAN DAN TUASAN TINGGI

Oleh

SYAIBA BALQISH BINTI ARIFFIN @ MAT ZIN

Disember 2018

Pengerusi : Profesor Habshah Midi, PhD
Fakulti : Sains

Model regresi binari logistik popular digunakan dalam analisis data perubatan. Di sebalik kepopularitiannya, hanya terdapat beberapa kaedah teguh bagi model ini untuk menangani kesan tuasan tinggi dan multikolinearan. Kegagalan untuk menangani kecukupan model bila kombinasi tuasan tinggi dan multikolinearan wujud dalam data, membawa kepada kekeliruan dan pentadbiran yang salah. Kajian ini bertujuan untuk membangunkan diagnosis dan penganggaran teguh yang baru untuk regresi logistik (kes-kes bertindih) dan regresi logistik tersembunyi (kes-kes tidak bertindih).

Satu diagnostic teguh yang baru dipanggil penama terpencil berpengaruh logistik (LION) dibangunkan untuk mengenalpasti terpencil berpengaruh dan LION berjaya mengesan terpencil dalam kedua-dua arah x dan y. Kemudian, diagnosis teguh kedua dinamakan diagnosis cerapan berpengaruh (DIO) dibangunkan, khusus untuk mengenalpasti tuasan tinggi cerapan berpengaruh (HLIO). DIO memperkenalkan dua peringkat penting di mana peringkat awal menggunakan prosedur LION dan peringkat pengesahan melibatkan gabungan ukuran pengitlakan jarak dari purata dan pengitlakan piawai Pearson reja untuk menandakan HLIO.

Ubahsuai pemberat Bianco dan Yohai (AWEBY) adalah penambahbaikan ke atas penggangar teguh pemberat Bianco and Yohai (WBY). AWEBY dicadangkan untuk meningkatkan kecekapan penganggar WBY dengan membina "penolakan licin" untuk menggantikan "penolakan keras" fungsi pemberat. Dalam AWEBY, pemberat teguh yang baru diformulasi berdasarkan

DIO dan didapati menurunkan secara tertib kesan HLIO juga memelihara tuasan tinggi baik. Dalam masalah gabungan HLIO and multikolinearan untuk kes-kes bertindih, penganggar AWEBY digabungkan untuk mengira parameter Ridge teguh dan membentuk logistik Ridge teguh (RRL) lelaran kemaskini skema. Dengan menggunakan pemberat teguh terkini, kesan HLIO dan multikolinearan akan diturunkan segera.

Ubahsuai pemberat maksimum kebarangkalian bolehanggar (AWEMEL) dalam regrasi logistik tersembunyi dicadangkan untuk membetulkan HLIO dalam kes-kes tidak bertindih. Pemberat tenguh yang baru dalam AWEMEL direka berdasarkan DIO yang terutamanya menurunkan HLIO tetapi bukan tuasan tinggi baik. Akhir sekali, logistik tersembunyi Ridge teguh (RRHL) dicadangkan untuk merawati kedua-dua HLIO dan multikolinearan untuk masalah kes-kes tak bertindih. Dalam lelaran RRHL, penggangar AWEMEL digunakan untuk mengira parameter Ridge teguh yang rintang terhadap kesan buruk HLIO.

ACKNOWLEDGEMENTS

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will remember and cherish the most. Foremost, I would like to express my deepest gratitude to my Supervisor, Professor Dr. Habshah Midi for her continuous support, motivation, enthusiasm, immense knowledge, caring, patience and providing me with an excellent atmosphere for doing research. It has been a great pleasure to work with and learn from such an extraordinary individual. Beside my Supervisor, I would like to thank the rest of my Supervisory Committee; Associate Professor Dr. Jayanthi Arasan and Dr. Md. Sohel Rana for their encouragement and insightful comments to better my work. I am particularly grateful to Professor Dr. A.H.M.R. Imon from Ball State University, Muncie, United State of America, for his comments and criticism on my work during his visits to Institute for Mathematical Research and as co-author of my papers. My sincere thanks also go to support staffs at Department of Mathematics, especially to Mr. Kamarulzaman Shah Bazil for providing me with a comfortable room to work and fast facilitate whenever I called out for help even during weekend. A million thanks to my scholarship providers; the Ministry of Higher Education Malaysia for MyBrain15MyPhD and the Universiti Putra Malaysia for Graduate Research Fellowship and Special Graduate Research Allowance during my study. Thanks are also due to many good friends and team of robust statistics that I have made in Universiti Putra Malaysia. In particular, I'd like to thank my doctoral friends; Dr. Nor Fadzillah Mohd Mokhtar, Dr. Aliyu Usman Moyi, Dr. Sirajo Bichi Lawan, Dr. Sarkhosh Seddighi Chaharborj, Dr. Mohammed Alguraibawi, Dr. Hassan Uraibi, Dr. Mohd Shafie Mustafa, Dr. Phang Pei See, Dr. Loh Yue Fang, Dr. Paul Dalatu Inuwa, Dr. Nor Mazlina Abu Bakar, Dr. Punitha Sinnapan, Dr. Syarifah Nasrisya Syed Nor Azlan, Dr. Khairul Nizam Samsudin, Rifina Arlin and late Saodah Ismail for the many fine times that we've had which provided a much needed diversion from our studies. Last but not least, I owe my deepest thanks to my parents; late Haji Ariffin Ismail, Hajah Sjarkiah Muhd Yasie, and my siblings; Malinda Ulfah, Azah Rahmi and Hambarjam, for their love, affection, encouragement and prayers for days and nights make me able to get such success and honor.

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

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

Jayanthi Arasan, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Md. Sohel Rana, PhD
Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
(Member)

_____
ROBIAH BINTI YUNUS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: _____ Date:_____

Name and Matric No.: Syaiba Balqish binti Ariffin @ Mat Zin (GS29371)

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to

Signature: _____
Name of Chairman of Supervisory Committee: Professor Dr. Habshah Midi

Signature: _____
Name of Member of Supervisory Committee: Associate Professor Dr. Jayanthi Arasan

Signature: _____
Name of Member of Supervisory Committee: Dr. Md. Sohel Rana

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AC	Outlier Detection Accuracy
AMSTLE	Adaptive Maximum Symmetrically Trimmed Likelihood Estimator
ASV	Asymptotic Variance Efficiency
AURL	Almost Unbiased Ridge Logistic
AWEBY	Adjusted Weighted Bianco and Yohai
AWEMEL	Adjusted Weighted Maximum Estimated Likelihood
BACON	Block Adaptive Computationally Efficient Outlier Nominator
BDHS	Bangladesh Demographic and Health Survey
BLP	Bad Leverage Point
BOFOLS	Best Omitted from the Ordinary Least Squares
BY	Bianco and Yohai
CB	Classical Bootstrap
CD	Cook's Distance
CE	Classification Error
CEO	Collinearity Enhancing Observation
CI	Condition Indexes
CIO	Collinearity Influential Observation
CME	Covariance Matrix Equality
CN	Condition Number
CRAN	Comprehensive R Archive Network
CRO	Collinearity Reducing Observation
C-Step	Concentration Step
CUBIF	Conditionally Unbiased Bounded Influence
DBB	Diagnostic Before Bootstrap
DFFITS	Different of Fits
DIO	Diagnostic Influential Observation
DM	Distance from the Mean
DPL	Double Penalized Log-likelihood
FMCD	Fast Minimum Covariance Determinant
FN	False Negative
FP	False Positive
FTL	Fast Trimmed Likelihood
GDFFITS	Generalized Different of Fits
GDM	Generalized Distance from the Mean
GLM	Generalized Linear Model
GLP	Good Leverage Point
GP	Generalized Potential
GSPR	Generalized Standardized Pearson Residual
GW	Generalized Weights
HLCEO	High Leverage Collinearity Enhancing Observation
HLCRO	High Leverage Collinearity Reducing Observation
HLIO	High Leverage Influential Observation
HLR	Hidden Logistic Regression
HLP	High Leverage Point
IO	Influential Observation
IRLS	Iterative Reweighted Least Squares
ISE	Index Set Equality

LION	Logistic Influential Outlier Nominator
LLP	Low Leverage Point
LMS	Least Median Squared
LO	Leverage Outlier
LTS	Least Trimmed Squared
MAD	Median Absolute Deviation
MCD	Minimum Covariance Determinant
MD	Mahalanobis Distance
MDDM	Median Deletion Distance from the Mean
ME	Misclassification Error
MEL	Maximum Estimated Likelihood
ML	Maximum Likelihood
MP1	M-estimator Pregibon 1
MP2	M-estimator Pregibon 2
MRCD	Minimum Regularized Covariance Determinant
MSE	Mean Squared Error
MSPR	Modified Standardized Pearson Residual
MSTLE	Maximum Symmetrically Trimmed Likelihood Estimator
MTL	Maximum Trimmed Likelihood
mvBACON	Multivariate Block Adaptive Computationally Efficient Outlier Nominator
MVE	Minimum Volume Ellipsoid
mvLION	Multivariate Logistic Influential Outlier Nominator
MVV	Minimum Vector Variance
OLS	Ordinary Least Square
PE	Prediction Error
PL	Penalized Log-Likelihood
regLION	Regression Logistic Influential Outlier Nominator
RLGD	Robust Logistic Diagnostic
RMD	Robust Mahalanobis Distance
RMVV	Reweighted Minimum Vector Variance
RO	Residual Outlier
RobDEVC	Robust Deviance Component
RHL	Ridge Hidden Logistic
RR	Ridge Regression
RRHL	Robust Ridge Hidden Logistic
RRL	Robust Ridge Logistic
SPR	Standardized Pearson Residual
SVD	Singular Value Decomposition
VDP	Variance Decomposition Proportion
VIF	Variance Inflation Factor
WBY	Weighted Bianco and Yohai
WBP	Weighted Bootstrap with Probability
WLGBP	Weighted Logistic Bootstrap with Probability
WEMEL	Weighted Maximum Estimated Likelihood
WML	Weighted Maximum Likelihood
WVIF	Weighted Variance Inflation Factor

CHAPTER 1

INTRODUCTION

1.1 Research Background

A logistic function was developed in the 19th century by Verhulst, who made an investigation of the population growth in the United State of America. The expression or word "logistic" was not applied until Verhulst's research was re-developed, which was immediately after Pearl and Reed first article that was published in the year 1920 (Cramer, 2002). The logistic function as an alternative to normal probability function is a research of Berkson (1944), who showed how the model could be fitted using iterative weighted least squares. Meanwhile, comparisons between logistic and probit transformation made by Chambers and Cox (1967) and Berkson (1951) which the difference mainly in the link function and distribution of the errors. Continuity of Berkson's work, Cox (1958a, 1958b) developed the logistic function for a binary response.

This thesis covers the case of binary response for logistic regression model where it can take only two values of outcome, such as success or failure, survive or dead, effective or impotent and etc. The binary logistic regression is a special case of the generalized linear model. This model measures a relationship between the binary response with both continuous and categorical predictor variables by estimating probabilities using the logistic function, which is the cumulative logistic distribution.

The binary logistic regression model, however, is based on quite different assumptions from those of linear regression model. In particular, the key differences between these two models can be seen in the following two features of binary logistic regression. First, the conditional mean $E(Y|X)$ is a Bernoulli distribution rather than a Gaussian distribution. Second, the predicted values are probabilities and the conditional mean lies between the ranges of $0 \leq E(Y|X) \leq 1$ where the change in $E(Y|X)$ per unit change in X become progressively smaller as the conditional mean gets closest to 0 or 1 (Hosmer and Lemeshow, 2000). It resembles a plot of a cumulative distribution of a random variable. Therefore, the binary logistic regression model can be graphically illustrated by a S-curve for one regressor and a hyperplane in the case of two regressors (Croux et al., 2002).

The parameters of binary logistic regression are usually estimated by the Maximum Likelihood (ML) estimator due to tradition and ease of computation (Albert and Anderson, 1984). However, the ML estimator is not robust against unusual observation (or known as an outlier) and it has a zero breakdown point

towards the outlier (Croux et al., 2002). Even a single outlier can drastically change the ML estimate very badly (Pregibon, 1981).

The ML estimator also becomes inefficient when there is a high degree of correlation among the continuous predictor variables that we refer to multicollinearity problem (Schaefer et al., 1984; Schaefer, 1986). Multicollinearity can result in a wrong sign problem and expand the magnitude of estimated regression coefficients, lead to erroneous interpretation and cause the estimated regression estimates have unduly large variances (Mackinnon and Puterman, 1989; Marx and Smith, 1990; Weissfeld and Sereika, 1991; Segerstedt and Nyquist, 1992; Lesaffre and Marx, 1993; Barker and Brown, 2001). Furthermore, Lesaffre and Marx (1993) remarked that inter-correlation problem in the binary model happens to occur in two situations i.e. correlation among predictor variables $(X'X)$ and correlation among weighted predictor variables $(X'WX)$ related to the Fisher information matrix.

Another problem arises from the binary logistic regression model is when the maximization of log-likelihood function fails to converge. In most cases, this failure is a consequence of data patterns known as quasi-complete separation (little overlap cases) and complete separation (non-overlap cases) (Albert and Anderson, 1984). The ML estimates simply do not exist in complete separation while quasi-complete separation can cause biased estimates with enormous standard errors. Separation in logistic regression frequently occurs when the binary response can be completely separated by a single predictor variable or by a linear combination of the predictor variables (Christmann and Rousseeuw, 2001b). Heinze and Schemper (2002) mentioned that the possibility of separation to occur is highly depends on the number of sample sizes, the number of dichotomous predictor variables and the magnitude of the odds ratios.

In the situations where finite ML estimates is not existing, Rousseeuw and Christmann (2003) introduced a Maximum Estimated Likelihood (MEL) as an alternative solution to the ML estimator. Rousseeuw and Christmann (2003) called this model as a Hidden Logistic Regression (HLR) model. It is evident that the MEL estimator is robust against separation and the MEL estimates always exist. Unfortunately, the MEL still has the disadvantage, that it is not robust to the presence of outliers and the influential observations because the impact of these observations is unbounded since the pseudo-observations, $\tilde{Y}_i$ is fitted to the MEL using the similar approach to the ML algorithm. Moreover, multicollinearity is also a common problem happen in the HLR model particularly when dealing with a small dataset.

In the next section, the shortcomings of classical diagnostic and estimation methods are addressed when handling with more complicated problem i.e. the combination of high leverage influential observation (HLIO) and multicollinearity which later becomes our motivations for this study.

1.2 Importance of the Study

Problem and Motivation One: This thesis concerned on the diagnostic method for identifying influential outliers in binary logistic regression. The influential outlier is defined as an observation that lies with abnormal distance from the majority of observations in covariate space and it has large residual. The influential outlier is strongly influencing the fitted equation, in such cases, both the intercept and slope of regression are severely affected. Therefore, the detection method is imperative because the influential outliers accountable for inaccurate prediction and invalid inferential statements as the influential outliers have a large impact on the computed values of various estimates. Literature is abundant with detection method for a single and multiple outlier in univariate data. To date, most of the diagnostic methods available for binary logistic regression model identifies outliers in x and y directions separately, which are not successful in detecting "genuine" influential outliers.

Billor et al. (2000) proposed Blocked Adaptive Computationally Efficient Outlier Nominator (BACON) for multivariate x-outliers and y-outliers in linear regression model. Meanwhile Müller and Neykov (2003) proposed Fast Trimmed Likelihood (FTL) for detecting y-outliers in Generalized Linear Model (GLM). Both methods exploit the concentration-steps (C-Step) iterative algorithm by Rousseeuw and Van Driessen (1999). The BACON method possesses good properties with affine equivariance, 50% high breakdown point, bounded influential function and surpasses minimum covariance determinant (MCD) method proposed by Rousseeuw and Van Driessen (1999) in term of fast convergence rate. The multivariate part (mvBACON) estimates the location and scatter matrix to compute Mahalanobis distances (MD), while regression part (regBACON) computes internal and external Studentized residuals derived from Ordinary Least Square (OLS) estimates. The BACON major drawback is that the mvBACON has to reimburse with low statistical efficiency and it gives bias estimates for a small data. Meanwhile, regBACON cannot be applied directly to binary logistic regression data due to different residuals measure. Thus, by applying only mvBACON to binary logistic regression data will not detect "genuine" influential outliers.

In GLM, Müller and Neykov (2003) proposed the FTL estimator where identification of regression y-outliers is based on log-likelihood measure. At first thought, it seems attractive to use the FTL in binary logistic regression. However, when applied trimming in C-steps, most probably the trimmed observations that are considered as y-outliers are the same observations that provide some overlap in the data. Therefore, trimming these observations eliminate the overlap cases, thus maximum log-likelihood value is undetermined (Christmann and Rousseeuw, 2001b).

The shortcomings of BACON and FTL have inspired us to develop a new diagnostic method to identify multivariate x-outliers and y-outliers in binary logistic regression model, namely Logistic Influential Outlier Nominator (LION)

Problem and Motivation Two: Depth investigation on leverage outlier (LO) or high leverage point (HLP), residual outlier (RO) and influential observation (IO) is important since outliers and IO are strongly related to one another but not interrelated. The first condition was mentioned by Chatterjee and Hadi (1988) where the IO need not to be outlying in the sense of having a large residual. Second, the inlying IO distort the shape of the fitted equation, though it has a small residual. Therefore, similar measures applied to detect HLP and RO are not relevant to detect IO.

Cook's Distance (CD) and Difference in Fits (DFFITS) are the commonly used methods for identifying IO. Nonetheless, Pregibon (1981) recommended using the DFFITS as it combines both the leverage and the residual components. Even though DFFITS is successful in detecting a single IO, it is not effective enough when multiple IOs are present in a data and the DFFITS becomes ineffective for identification of multiple IOs due to masking and swamping effects (Nurunnabi et al., 2010).

Nurunnabi et al. (2010) developed the generalized version of DFFITS, denoted as GDFFITS which combined both measures of the group deleted leverage and residual components. Although the GDFFITS can detect multiple IOs, it is not effective enough in identifying the exact number of IO. It has a tendency of detecting lesser IO as it should be and produce several masking IOs. This is most probably due to the determination of the initial subset of the GDFFITS which is not adequately effective in classifying the deletion and the remaining groups. The GDFFITS exploits the BACON method to identify a set of suspected IOs, denoted as set D. Although the expression for GDFFITS is available for any arbitrary set of deleted cased D, the choice of such a set is the most important component in diagnostic procedure since the omission of set D determines the correct GDFFITS values for both set D and the remaining set R. As we already mentioned, the regression BACON (regBACON) is inadmissible for binary response and by only depending on multivariate BACON (mvBACON), there is a high possibility for suspected IO remains in the set R, thus the entire GDFFITS value is wrongly calculate. Moreover, no further discussion on whether the detection of IO by GDFFITS classified a good or bad IO. However, the GDFFITS method has inspired us to classify IO into good and bad IO.

The weakness of Nurunnabi et al. (2010) approach has motivated us to propose a new robust diagnostic method, namely Diagnostic Influential Observations (DIO) whereby the suspected high leverage influential observations (HLIO) in set D are identified using the LION method.

Problem and Motivation Three: Abundant of studies has been carried out in developing robust estimators without paying much attention whether or not influence function of robust estimators were bounded to good leverage point (GLP) or bad leverage points (BLP). It is imperative to distinguish between the GLP and BLP as the BLP extremely influential to fitted model while not the GLP.

In this regard, we take initiative to improve the Weighted Bianco and Yohai (WBY) estimator proposed by Croux and Haesbroeck (2003). The WBY applies Robust Mahalanobis Distance (RMD) based on MCD estimator which is less efficient as it downweight all detected HLPs irrespective of whether it is GLP or BLP. Moreover, Croux and Haesbroeck (2003) implemented hard rejection weight based RMD-MCD to reduce the effect of HLPs. Referring to this approach, HLPs were assigned to zero weight and excluded before estimation. It is evident that deleting the GLPs reduces the precision of estimates and increase the possibility for cases to be separated (Croux, 2006).

In this situation, we proposed an Adjusted Weighted Bianco and Yohai (AWEBY) estimator with a modification of weighting scheme. A smooth rejection weight is formulated based on the DIO values. By applying the new weight, elimination is restricted to the BLPs while the GLPs are protected, thus significantly improves the precision of AWEBY estimates.

Problem and Motivation Four: Montgomery and Peck (1982) and Gunst (1983) pointed out that there are different sources of multicollinearity such as data collection, method employed constraint on the model, model specification and over determined model. To remedy multicollinearity problem which is due to these sources, Mansson and Shukur (2011) proposed using Ridge Logistic (RL) estimators. Nevertheless, Mansson and Shukur (2011) did not discuss any method of how to rectify the problems when both multicollinearity and outliers occur together in data. The presence of outlier in multicollinear data creates misleading conclusion on RL estimates.

Since not much research has been done in exploiting these issues, this motivates us to come up with a new Robust Ridge Logistic (RRL) estimator for binary logistic regression model. The RRL incorporates the AWEBY estimator to obtain robust ridge parameter towards the BLPs which later is used in ridge iterative update scheme to reduce the variance inflation due to multicollinearity. We expect that the newly developed method would be more efficient than the existing RL estimates, since we would remove the influence of outliers and multicollinearity problems by the robust AWEBY estimator which will be embedded in the RRL estimator.

Problem and Motivation Five: Rousseeuw and Christmann (2003) proposed Weighted Maximum Estimated Likelihood (WEMEL) estimator to rectify the problem of HLPs and separation cases. It is evident that the WEMEL estimator

is resistant against separation, bounded to the HLPs and the estimates always exist. Unfortunately, the WEMEL uses the similar approach with the WBY in treating HLPs where non-hard rejection weight is computed based on RMD-MCD.

The weakness of RMD-MCD is that it is prone to swamping effect where some of GLPs are detected as BLPs. Thus, decrease the efficiency of the WEMEL estimates. This inspired us to propose an Adjusted Weighted Maximum Estimated Likelihood (AWEMEL) whereby the GLPs are not downweighted.

Problem and Motivation Six: Studies that investigate the multicollinearity in separation problem is still in infant stage. The most current method which provides a solution to both problems of separation and multicollinearity is a Double Penalized Likelihood Estimator (DPL) proposed by Shen and Gao (2008). The DPL method applied Jeffrey's non-informative prior and the ridge type method and the computation of ridge parameter is by a cross validation process which minimizes the mean squared error (MSE) of DPL estimate. Godínez-Jaimes et al. (2012) claimed that the DPL estimator is not actually remedy the effect of multicollinearity in separation data. According to simulation experiment conducted by Godínez-Jaimes et al. (2012), ridge logistic type estimator proposed by Le Cessie and Van Houwelingen (1992) defeats the DPL estimator, evident by lower biases and MSEs. Moreover, there is a lack of literature dealing with simultaneous problem of outliers and multicollinearity for separation cases in binary logistic regression model.

This has motivated us to investigate such complex scenario whereby both collinear and separated data simultaneously occurs in the presence of HLPs. In order to remedy these problems, we proposed a Robust Ridge Hidden Logistic (RRHL) by incorporating the AWEMEL estimator to compute robust ridge parameter for variance reduction of highly correlated predictors while handling the outlier-separation issues.

1.3 Research Objectives

The major purpose of this thesis is to investigate the effect of HLIO on the parameter estimation of binary logistic regression model for overlap and non-overlap cases. Then, we extend the work of investigating a combined problem of multicollinearity and HLIOs for overlap and non-overlap cases. The current diagnostic procedures and robust estimation methods deal with one type of outlier at a time. Thus, the development of proposed robust diagnostic and estimation methods are crucial. The foremost objectives of our research can be outlined systematically as follows:

1. To develop a new robust diagnostic procedure for identifying influential outliers in multivariate data.

2. To develop a new robust diagnostic procedure for identifying high leverage influential observations.
3. To propose a new robust estimator in the presence of high leverage influential observations when cases are overlapping.
4. To propose a new robust ridge estimator having both multicollinearity and high leverage influential observations when cases are overlapping.
5. To propose a new robust estimator in the presence of high leverage influential observations when cases are separating.
6. To propose a new robust ridge estimator having both multicollinearity and high leverage influential observations when cases are separating.

1.4 Scope and Limitation of the Study

The robust diagnostic procedures and robust estimation methods for this model is limited to binary response with continuous predictor variables. For a comprehensive evaluation of the proposed methods, several factors are investigated which include the number of observations, n, number of continuous predictor variables, p, number of contaminations, e, and the degree of correlations, ρ.

In this thesis, we consider a matrix X with dimension $n \times p$ where $p < n$. For data with overlap cases, simulation experiments vary with sample sizes within range of $100 \leq n \leq 1000$ while data with non-overlap cases start with smaller sample size $20 \leq n \leq 1000$. The continuous predictor variables are set as $2 \leq p \leq 10$. The percentages of contaminations plugged in data with $1\% \leq e \leq 10\%$ out of n where good observations are replaced with contamination values and the degree of correlations varies within range $0.75 \leq \rho \leq 0.99$.

We use benchmark datasets for the identification of outliers and multicollinearity in binary logistic regression. However, some of the real datasets are not within the scope of simulation studies, particularly for overlapping cases. It is difficult to obtain dataset which has a combined problem of multicolinearity and HLPs in overlap and non-overlap cases, since not much work have been focused to deal with these problems. Therefore, for a small real dataset and small random generated data which prone to have separation cases, we suggest to apply the RRHL-AWEMEL estimators.

1.5 Outline of the Thesis

This thesis pursues with the newest robust diagnostic procedures and the new robust estimation methods for overlap and non-overlap cases in the presence of both HLIOs and multicollinearity for logistic regression model with binary response. Our contribution chapters begin from Chapter 3 to Chapter 8. The

new proposed methods are extensively investigated by simulation studies (due to little theoretical justification on proposed methods) and application on datasets available from literatures. The remainder of this thesis is organized as follows.

Chapter Two: This chapter reviewed on the estimation methods for logistic regression (overlap cases) and hidden logistic regression (non-overlap cases) and violation of model assumptions due to the presence of outliers, influential observations and multicollinearity. Detailed definitions of various types of outliers, current diagnostic procedures and effect of outliers on parameter estimations are also discussed. The review on multicollinearity covered the types of ill-conditioning, sources and effects, current diagnostic measures and ridge regression estimation. Furthermore, the effect of collinearity influential observations on multicollinearity diagnostic procedure is also highlighted. Finally, basic concepts of good robust estimation and some important existing robust estimators are also included.

Chapter Three: The development of a new proposed LION method is shown in this chapter. The LION is designed to identify the influential outliers where detection procedure tackles both x and y outliers.

Chapter Four: A new diagnostic method to identify HLIO is proposed in this chapter. The development of DIO method consists of two stages. The initial stage employs the LION procedure and the confirmation stage comprises a combine measures of GDM and GSPR to flag the HLIO.

Chapter Five: In this chapter, the WBY estimator is improved by constructing a smooth rejection weight function to replace a hard rejection one. In the development of a new AWEBY estimator, the DIO is incorporated to formulate the new weight, which properly reduce the effect of the HLIO while protecting the GLP for more precise AWEBY estimates.

Chapter Six: This chapter deals with the development of a new RRL estimator for a combined problem of multicollinearity and HLIO in overlap cases. In the RRL iterative update scheme, the AWEBY estimator is integrated in computing robust ridge parameter which plays an important role to handle the HLIO-multicollinearities dataset.

Chapter Seven: In this chapter, a new AWEMEL estimator in HLR is proposed to rectify the HLIO in separation problem. The WEMEL is a good estimator despite of its estimation is less precise compared to AWEMEL. A new weight in the AWEMEL is designed based on DIO where only BLP are downweighted instead of HLP.

Chapter Eight: This chapter involves proposing a new estimator, the RRHL to remedy both multicollinearity and HLIO for separation problem. In ridge's iteration, the AWEMEL estimator is employed to compute robust ridge parameter which resistance to the HLIO.

Chapter Nine: This chapter provides summaries and conclusions on proposed methods. Areas for further research are also discussed.

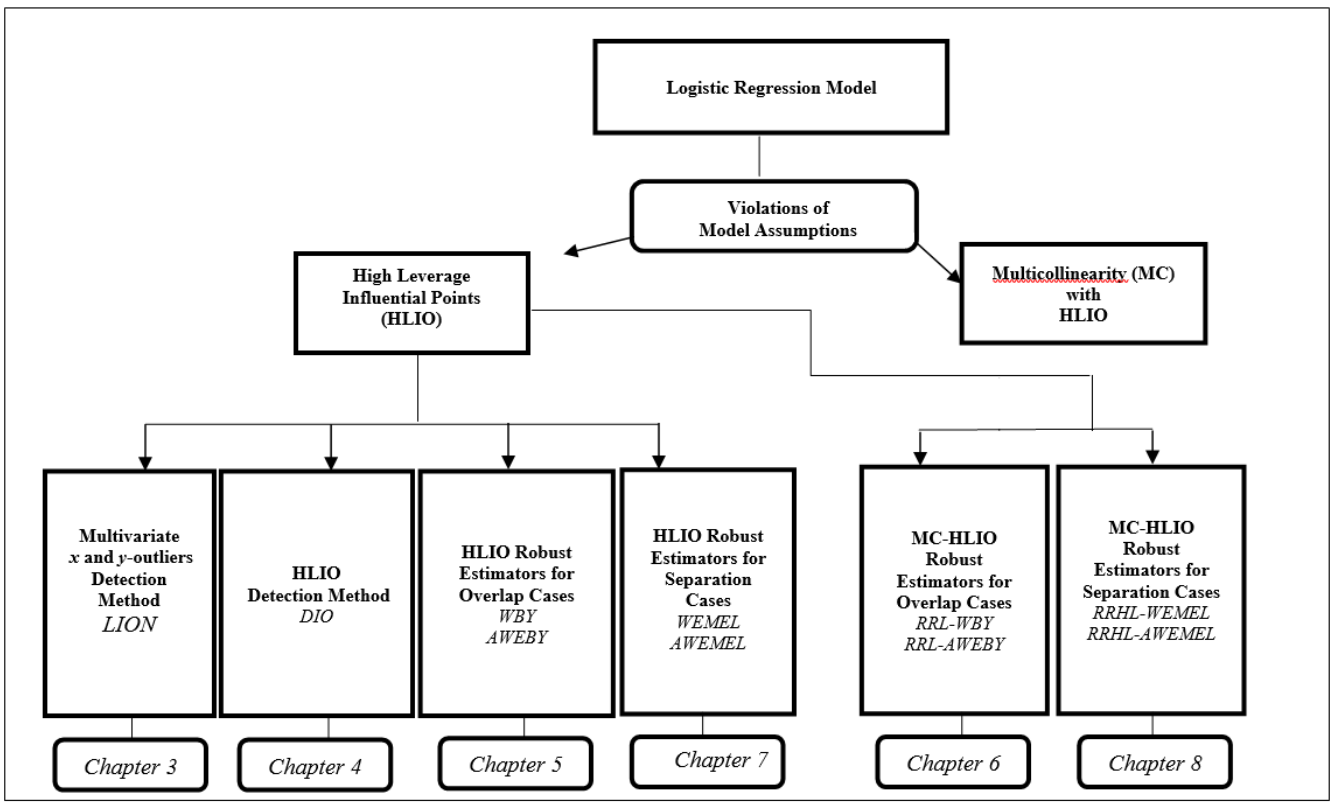

Figure 1.1: Flowchart of the study

REFERENCES

Adimari, G., and Ventura, L. (2001). Robust inference for generalized linear models with application to logistic regression. *Statistics and Probability Letters*. 55(4):413-419.

Ahmad, S., Ramli, N.M., and Midi, H. (2010). Robust estimators in logistic regression: A comparative simulation study. *Journal of Modern Applied Statistical Methods*. 9(2):502-511.

Aitkin, M., Anderson, D.A., Francis, B., and Hinde, JP. (1989). *Statistical modelling in GLIM 4*. Oxford:Clarendon Press.

Al-Aabdi, F.A.A., and Al–Shaibani, R.M.A. (2014). Robust estimators of logistic regression with problems multicollinearity or outliers values. *Journal of Kufa for Mathematics and Computer*. 2(2):64-71.

Albert, A., and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression. *Biometrika*. 71(1):1-10.

Ali, H., Syed Yahaya, S.S., and Omar, Z. (2014). The efficiency of reweighted minimum vector variance. AIP Conference Proceedings. 1602(1):1151-1156.

Amado, C., and Pires, A.M. (2004) Robust bootstrap with non-random weights based on the influence function, *Communications in Statistics - Simulation and Computation*, 33(2) :377-396. Analysis, 48(4), 703-715.

Amemiya, Y. (1985). What should be done when an estimated between-group covariance matrix is not nonnegative definite? *The American Statistician*. 39(2):112-117.

Andrews, D.F. (1971). A note on the selection of data transformations. *Biometrika*. 58(2):249-254.

Anscombe, F.J. (1960). Rejection of outliers. *Technometrics*. 2(2):123-146.

Ariffin, S.B., and Midi, H. (2010). Robust logistic diagnostic for the identification of high leverage points in logistic regression model. *Journal of Applied Sciences*. 10(23):3042-3050.

Babaie-Kafaki, S., and Roozbeh, M. (2017). A revised Cholesky decomposition to combat multicollinearity in multiple regression models. *Journal of Statistical Computation and Simulation*, 87(12): 2298-2308.

Bagheri, A., and Midi, H. (2012). On the performance of the measure for diagnosing multiple high leverage collinearity-reducing observations. *Mathematical Problems in Engineering*, Volume 2012, Article ID 531607, 16 pages.

Bagheri, A., Habshah, M., and Imon, A.H.M.R. (2012). A novel collinearity-influential observation diagnostic measure based on a group deletion approach. *Communications in Statistics-Simulation and Computation*. 41(8):1379-1396.

Bagheri, A., and Midi, H. (2015). Diagnostic plot for the identification of high leverage collinearity-influential observations. *SORT-Statistics and Operations Research Transactions*. 39(1):51-70.

Barker, L., and Brown, C. (2001). Logistic regression when binary predictor variables are highly correlated. *Statistics in Medicine*. 20(9-10):1431-1442.

Barnett, V., and Lewis, T. (1994). *Outliers in statistical data* 3rd Ed. Chichester:Wiley.

Barreto, I.D.D.C., Russo, S.L., Brasil, G.H., and Simon, V.H. (2014). Separation Phenomena Logistic Regression. *Revista GEINTEC-Gestão, Inovação e Tecnologias*. 4(1): 716-728.

Bedrick, E.J., and Hill, J.R. (1990). Outlier tests for logistic regression: a conditional approach. *Biometrika*. 77(4):815-827.

Begg, M.D., and Lagakos, S. (1990). On the consequences of model misspecification in logistic regression. *Environmental Health Perspectives*. 87:69-75.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York:Wiley.

Belsley, D.A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*. 38(2):73-77.

Belsley, D.A., and Oldford, R.W. (1986). The general problem of ill conditioning and its role in statistical analysis. *Computational Statistics and Data Analysis*. 4(2):103-120.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*. 39(227):357-365.

Berkson, J. (1951). Why I prefer logits to probits. *Biometrics*. 7(4):327-339.

Bianco, A.M., and Yohai, V.J. (1996). Robust estimation in the logistic regression model. Robust statistics, data analysis and computer intensive methods. Proceedings of the workshop in honor of Huber, P.J., and Rieder, H. Lecturer Notes in Statistics. 109:17-34. New York:Springer.

Bianco, A. M., and Martínez, E. (2009). Robust testing in the logistic regression model. *Computational Statistics and Data Analysis*, 53(12): 4095-4105.

Bickel, P., and Lehmann, E. (1976). Descriptive statistics for nonparametric problems. IV: Spread. *Contributions to Statistics, Juneckova (ed)*, 33-40.

Billor, N., Hadi, A.S., and Velleman, P.F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis*. 34(3):279-298.

Bondell, H.D. (2005). Minimum distance estimation for the logistic regression model. *Biometrika*. 92(3):724-731.

Boudt, K., Rousseeuw, P. and Vanduffel, S. and Verdonck, T. (2018). The Minimum Regularized Covariance Determinant. Pp.23.https://ssrn.com/abstract=2905259http://dx.doi.org/10.2139/ssrn.2905259.

Breslow, N.E., and Day, N.E. (1980). Statistical methods in cancer research. The analysis of case-control studies, 1(32) pages. Geneva:IARC Press.

Brown, B.W. (1980). *Prediction analysis for binary data*. In biostatistics casebook (pp.3-18). New York:Wiley.

Butler, R.W., Davies, P.L., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*. 21(3):1385-1400.

Carroll, R.J., and Pederson, S. (1993). On robustness in the logistic regression model. *Journal of Royal Statistics Society B*. 55(3):693-706.

Chambers, E.A., and Cox, D.R. (1967). Discrimination between alternative binary response models. *Biometrika*. 54(3-4):573-578.

Chatterjee, S., and Price B. (1977). *Regression analysis by example*. New York:Wiley.

Chatterjee, S., and Hadi, A.S. (1988). *Sensitivity analysis in linear regression*. New York:Wiley.

Christmann, A. (1994). Least median of weighted squares in logistic regression with large strata. *Biometrika*. 81(2):413-417.

Christmann, A., and Rousseeuw, P.J. (2001a). Measuring overlap in binary regression. *Computational Statistics and Data Analysis*. 37(1):65-75.

Christmann, A., and Rousseeuw, P.J. (2001b). The hidden logistic regression model. Technical Report (pp.1-14), Universitätsbibliothek Dortmund, Germany.

Čížek, P. (2006a). Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference*. 136(11):3967-3988.

Čížek, P. (2006b). Trimmed likelihood-based estimation in binary regression models. *Austrian Journal of Statistics*. 35(2-3): 223-232.

Čížek, P. (2008). Robust and efficient adaptive estimation of binary-choice regression models. *Journal of the American Statistical Association*. 103(482):687-696

Collett D. (2003). *Modelling binary data* 2nd Ed. (pp.65-71). London:Chapman and Hall.

Cook, R.D., and Hawkins, D. M. (1990). Comment on unmasking multivariate outliers and leverage points. *Journal of American Statistical Association*. 85(411):640-644.

Copas, J.B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society B*. 50(2):225-265.

Cox, D.R. (1958a). Two further applications of a model for binary regression. *Biometrika*. 45(3/4):562–565.

Cox, D.R. (1958b). The regression analysis of binary sequences. *Journal of the Royal Statistical Society B*. 20(2):215–242.

Cox, D.R.(1970). *Analysis of binary data*. London:Chapman and Hall.

Cramer, J.S. (2002). The origins of logistic regression. Discussion paper (No. 119/4). Tinbergen Institute.

Croux, C., and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*. 71(2):161-190.

Croux, C., Flandre, C., and Haesbroeck, G. (2002). The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Statistics and Probability Letters*. 60(4):377-386.

Croux, C., and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*. 44(1-2):273-295.

Croux C. (2006). Are good leverage points good or bad? Proceeding of the International Conference on Robust Statistics, Lisbon, Portugal, July 16-21.

Cule, E., and De Iorio, M. (2012). A semi-automatic method to guide the choice of ridge parameter in ridge regression. *The Annals of Applied Statistics*. 1(2):302-332.

Davies, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics*. 20(4):1828-1843.

Davies, L., and Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*. 88(423):782-792.

Davies, P.L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*. 15(3):1269-1292.

De Jongh, P.J., De Jonghy, E., Pienaarz, M., Gordon-Grantz, H., Oberholzerz, M., and Santana, L. (2015). The impact of pre-selected variance inflation factor thresholds on the stability and predictive power of logistic regression models in credit scoring. *ORiON*. 31(1):17-37.

De La Rey, T. (2007). Two statistical problems related to credit scoring Published Doctoral Dissertation, North-West University, Potchefstroom, South Africa.

Dehon, C., Gassner, M. and Verardi, V. (2009). Beware of good outliers and overoptimistic conclusions. *Oxford Bulletin of Economics and Statistics*. 71(3):437-452.

Djauhari, M. A., Mashuri, M., and Herwindiati, D. E. (2008). Multivariate process variability monitoring. *Communications in Statistics-Theory and Methods*, 37(11):1742-1754.

Donoho, D.L., and Huber., P.J. (1983). *The notion of breakdown point*. In a Festschrift for Lehmann, E. (pp.157-184). Belmont:Wadsworth Publishing.

Dorsett, D., Gunst, R.F., and Gartland, E.C. (1983). Multicollinear effects of weighted least squares regression. *Statistics and Probability Letters*. 1(4):207-211.

Efron, B. (1979) Bootstrap method: Another look at the Jackknife. *The Annals of Statistics*, 7(1) :1-26.

Efron, B. and Tibshirani, R.J. (1998) *An Introduction to the Bootstrap*, 105-120 p, Chapman and Hall: Boca Raton.

Ekholm, A., and Palmgren, J. (1982). *A model for a binary response with misclassifications in GLIM 82*. Proceedings of the International Conference on Generalised Linear Models (pp.128-143). New York:Springer.

Everitt, B.S. (1992). *The analysis of contingency tables* 2nd Ed. London:Chapman and Hall.

Fauconnier, C., and Haesbroeck, G. (2009). Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology*, 6(4):363-379.

Finney, D.J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika*. 34(3/4):320-334.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52(3): 1694-1711.

Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika*. 80(1): 27-38.

Fisher, R.A. (1925). *Theory of statistical estimation*. Mathematical Proceedings of the Cambridge Philosophical Society. 22(5):700-725. New York:Cambridge University Press.

Giaimo, R., Matranga, D., and Campisi, G. (2006). Odds ratio estimation in the presence of complete or quasi-complete separation in data. *Statistica Applicata*, 18(3):429-444.

Gervini, D. (2005). Robust adaptive estimators for binary regression models. *Journal of Statistical Planning and Inference*. 131(2):297-311.

Godinez-Jaimes, F., Ramirez-Valverde, G., Reyes-Carreto, R., Ariza-Hernandez, F. J., and Barrera-Rodriguez, E. (2012). Collinearity and Separated Data in the Logistic Regression Model. *Agrociencia*. 46(4):411-425.

Greenland, S., Schwartzbaum, J. A., and Finkle, W.D. (2000). Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology*. 151(5):531-539.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*. 11(1):1-21.

Gunst, R.F. (1983). Regresion analysis with multicollinear predictor variables. *Communications in Statistics-Theory and Methods*. 12(19):2217-2260.

Habshah, M., 2000, Bootstrap methods: A class of non-linear regression models. *Pertanika Journal of Science and Technology*, **8**(2), 175-189.

Habshah, M., Hassan, S.U. and Bashar, A.T. (2009). Dynamic robust bootstrap method based on LTS estimators, *European Journal of Scientific Research*, 32(3): 277-287.

Hadi, A.S. (1988). Diagnosing collinearity-influential observations. *Computational Statistics and Data Analysis*. 7(2):143-159.

Hadi, A.S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics and Data Analysis*. 14(1):1-27.

Hadi, A.S., and Simonoff, J.S. (1993). Procedure for the identification of outliers in linear models. *Journal of American Statistical Association*. 88(424):1264-1272.

Hampel, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics*. 27(2):95-107.

Hampel, F.R. (1971). A general definition of qualitative robustness. *The Annals of Mathematical Statistics*. 42:1887-1896.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*. 69(346):383-393.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust statistics: the approach based on influence functions* (pp.81-95). New York:Wiley.

Hao, Y. (1992). Maximum median likelihood and maximum trimmed likelihood estimations. Published Doctoral Dissertation, University of Toronto, Canada.

Haque, A., Jawad, A.F., and Cnaan, A. and Shabbout M. (2002). Detecting multicollinearity in logistic regression models: An extension of BKW diagnostic. Proceedings of Joint Statistical Meeting, American Statistical Association, New York. (pp. 1356-58).

Hawkins, D.M. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics and Data Analysis*. 17(2):197-210.

Heinze, G., and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*. 21(16):2409-2419.

Hermans., J., and Habbema, J.D.F. (1975). Comparison of five methods to estimate a prosteriori probabilities. *EDV Med. Biol*. 6: 14-19.

Herwindiati, D.E., Djauhari, M.A., and Mashuri, M. (2007). Robust multivariate outlier labeling. Communications in Statistics: *Simulation and Computation*. 36(6):1287-1294.

Hoaglin, D.C., and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *Journal of American Statistical Association*. 32(465):17-22.

Hobza, T., Pardo, L., and Vajda, I. (2008). Robust median estimator in logistic regression. *Journal of Statistical Planning and Inference*. 138(12):3822-3840.

Hoerl, A.E., and Kennard, R.W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 12(1):55-67.

Hoerl, A.E., and Kennard, R.W. (1970b). Ridge regression: applications to nonorthogonal problems. *Technometrics*. 12(1):69-82.

Hoerl, A.E., Kennard, R.W., and Baldwin, K.F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*. 4(2):105-123.

Hosmer, D.W. and S. Lemeshow, (2000). *Applied logistic regression* 2nd Ed. (pp.143-199). New York:Wiley.

Hossain, A., Khan, H.T.A. (2004). Nonparametric bootstrapping for multiple logistic regression model using R. *BRAC University Journal*, 1(2) :109-113.

Huber, P.J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*. 35(1):73-101.

Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*. 1(5):799-821.

Huber, P.J. (1981). *Robust statistics*. New York:Wiley.

Hubert, M. (1997). The breakdown value of the L1 estimator in contingency tables. *Statistics and Probability Letters*. 33(4):419-425.

Imon, A.H.M.R., Ali, M.M. (2005) Bootstrapping regression residual. *Journal of Korean Data and Information Science Society*, 16(3): 665-682.

Imon, A.H.M.R. (2005). Identifying multiple influential observations in linear regression. *Journal Applied Statistics*. 32(9):929-946.

Imon, A.H.M.R. (2006). Identification of high leverage points in logistic regression. *Pakistan Journal of Statistics*. 22(2): 147-156.

Imon, A.H.M.R., and Apu, M.R. (2007). Identification of multiple high leverage points using robust mahalanobis distance. *Journal of Statistics Studies*. 32:929-946.

Imon, A.H.M.R., and Hadi, A.S. (2008). Identification of multiple outliers in logistic regression. *Communications in Statistics-Theory and Methods*. 37(11):1697-1709.

Imon, A.H.M.R., and Hadi, A.S. (2013). Identification of multiple high leverage points in logistic regression. *Journal of Applied Statistics*. 40(12):2601-2616.

Ishwaran, H. (1999). Applications of hybrid Monte Carlo to Bayesian generalized linear models: Quasi-complete separation and neural networks. *Journal of Computational and Graphical Statistics*. 8(4):779-799.

Izrael, D., Battaglia, A.A., Hoaglin, D.C., Battaglia, M.P. (2002). Use of the ROC curve and the bootstrap in computing weighted logistic regression models, 248-270 p, In Proceedings of the Twenty-seventh Annual SAS Users Group International Conference, Cary, Sugi27, Statistics and Data Analysis.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A*, 186(1007):453-461.

Jennings, D.E. (1986). Outliers and residual distributions in logistic regression. *Journal of the American Statistical Association*. 81(396):987-990.

Kibria, B.M.G., Månsson, K., and Shukur, G. (2012). Performance of some logistic ridge regression estimators. *Computational Economics*. 40(4):401-414.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982). *Epidemiologic research: principles and quantitative methods*. New York:Wiley.

Kondylis, A., and Hadi, A. S. (2006). Derived components regression using the BACON algorithm. *Computational Statistics and Data analysis*, 51(2):556-569.

Konis, K. (2007). Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models. Published Doctoral Dissertation, Worcester College University of Oxford.

Kordzakhia, N., Mishra, G.D., and Reiersølmoen, L. (2001). Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference*. 98(1):211-223.

Künsch, H.R., Stefanski, L.A., and Carroll, R.J. (1989). Conditionally unbiased bounded influence estimation in general regression models with applications to generalized linear models. *Journal of American Statistical Association*. 84(46):460-466.

Lawrence, K.D, and Arthur, J.L. (1990). *Robust regression: Analysis and application*. New York: Marcel Dekker.

Le Cessie, S., and Van Houwelingen, J.C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society C*. 41(1):191-201.

Lesaffre, E. and Marx, B.D. (1993). Collinearity in generalized linear regression. *Communications in Statistics-Theory and Methods*. 22(7):1933-1952.

Linhart, H., and Zucchini, W. (1986). Finite sample selection criteria for multinomial models. *Statistische Hefte*, 27(1), 173-178.

Lopuhaa, H.P., and Rousseeuw, P.J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*. 19(1):229-248.

Lopuhaa, H.P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *Annals of Statistics*. 27(5):1638-1665.

Mackinnon, M.J., and Puterman, M.L. (1989). Collinearity in generalized linear models. *Communications in statistics-theory and methods*. 18(9):3463-3472.

Månsson, K., and Shukur, G. (2011). On ridge parameters in logistic regression. *Communications in Statistics-Theory and Methods*. 40(18):3366-3381.

Markatou, M., Basu, A., and Lindsay, B. (1997). Weighted likelihood estimating equations: The discrete case with applications to logistic regression. *Journal of Statistical Planning and Inference*. 57(2):215-232.

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*. 12(3):591-612.

Marrona, R.A., Martin, R.D., and Yohai, V.J. (2006). *Robust statistics: Theory and Methods*. New York:Wiley.

Marx, B.D. (1989). *Ill-conditioned information matrices and the generalized linear model: an asymptotically biased estimation approach. In Statistical Modelling* (pp. 206-213). New York:Springer.

Marx, B.D., and Smith, E.P. (1990). Weighted multicollinearity in logistic regression: diagnostics and biased estimation techniques with an example from lake acidification. *Canadian Journal of Fisheries and Aquatic Sciences*. 47(6):1128-1135.

Mason, R.L., and Gunst, R.F. (1985). Outlier-induced collinearities. *Technometrics*. 27(4):401-407.

McCullagh, P., and Nelder, J.A. (1989). *Generalized linear models*, 2nd Ed. (pp.98-148). London:Chapman and Hall.

Mela, C.F., and Kopalle, P.K. (2002). The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. *Applied Economics*. 34(6):667-677.

Menard, S. (2002). *Applied logistic regression analysis* 2nd Ed. London:Sage

Meyer, B.D., and Mittag, N. (2017). Misclassification in binary choice models. *Journal of Econometrics*, 200(2):295-311.

Michalek, J.E., and Tripathi, R.C. (1980). The effect of errors in diagnosis and measurement on the estimation of the probability of an event. *Journal of the American Statistical Association*. 75(371):713-721.

Midi, H., Rana, Md. S. and Mat Said, N.A (2011). The application of two stage robust weighted least squares and robust bootstrapping procedure on food expenditure data. *International Journal of Applied Mathematics and Statistics*, 20(11) :25-37.

Midi, H., and Ariffin, S.B. (2012). The performance of classical and robust logistic regression estimators in the presence of outliers. *Pertanika Journal of Science and Technology*. 20(2):313-325.

Midi, H., and Ariffin, S. B. (2013). Modified Standardized Pearson Residual for the Identification of Outliers in Logistic Regression Model. *Journal of Applied Sciences*, 13(6), 828-836.

Midi, H., Ramli, N.M, and Imon, A.H.M.R. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*. 36(5):507-520.

Midi, H., Sarkar, S. K., and Rana, M.S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*. 13(3):253-267.

Montgomery, D.C. and Peck, E.A (1982). *Introduction to linear regression analysis*. New York:Wiley.

Müller, C.H., and Neykov, N. (2003). Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and inference*. 116(2):503-519.

Nelder J.A., and Wedderburn R.W.M.(1972). Generalized linear model. *Journal of the Royal Statistical Society A*. 135(3):370-384

Nguyen, T.D., and Welsch, R.E. (2010). Outlier detection and robust covariance estimation using mathematical programming. *Advances in Data Analysis and Classification*. 4(4):301-334.

Ramli, N.M., Ahmad, S., and Midi, H. (2011) Identifying bad leverage points in logistic regression model based on robust deviance components. Proceedings of the Mathematical Models and Methods in Modern Science (pp.62-67).

Norazan, M.R. (2008). Weighted Maximum Median Likelihood Estimation for Parameters in Multiple Linear Regression Model. PhD Thesis, Faculty of Science, Universiti Putra Malaysia, Malaysia, 269-332 p.

Norazan, M.R., Habshah, M., Imon, A.H.M.R. (2009) Estimating regression coefficients using weighted bootstrap with probability. *WSEAS Transactions on Mathematics*, 8(7): 362-371.

Nurunnabi, A.A.M., Imon, A.H.M.R., and Nasser, M. (2010). Identification of multiple influential observations in logistic regression. *Journal of Applied Statistics*. 37(10):1605-1624.

Park, H., and Konishi, S. (2016). Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. *Journal of Statistical Computation and Simulation*. 86(7):1450-1461.

Park, M.Y., and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*. 9(1):30-50.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., and Feinstein, A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*. 49(12):1373-1379.

Pierce, D.A., and Schafer, D.W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*. 81(396):977-986.

Pison, G., Van Aelst, S., and Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika*. 55(1-2):111-123.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics*. 29(1):15-23.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*. 9(4):705-724.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*. 38(2):485-498.

Roberts, G., Binder, D., Kovačević, M., Pantel, M., Phillips, O. (2003) Using an estimating function bootstrap approach for obtaining variances estimates when modelling complex health survey data, 1-9 p, In Proceeding of the Survey Methods Sections, SSC Annual Meeting.

Rocke, D.M., and Woodruff, D.L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*. 91(435):1047-1061.

Roelant, E., Van Aelst, S., and Willems, G. (2009). The minimum weighted covariance determinant estimator. *Metrika*. 70(2):177-204.

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*. 37(8):283-297.

Rousseeuw, P.J., and Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Journal of Computational Statistics and Data Analysis*. 43(3):315-332.

Rousseeuw, P.J., and Van Driessen K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*. 41(3):212-223.

Rousseuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P.J., and Hubert, M. (1999). Regression Depth. *Journal of the American Statistical Association*. 94(446):388–433.

Ryan, T.P. (2008). *Modern regression methods* (pp.255-313). New York:Wiley.

Sakata, S., and White, H. (1995). An alternative definition of finite-sample breakdown point with applications to regression model estimators. *Journal of the American Statistical Association*. 90(431):1099-1106.

Salibian-Barrera, M. (2006) Bootstrapping MM-estimators for linear regression with fixed designs. *Statistics and Probability Letter*, 76(12) :1287-1297.

Salibian-Barrera, M., Zamar, R.H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics*, 30(2) : 556-582.

Salleh R.M. (2013). A Robust Estimation Method of Location and Scale with Application in Monitoring Process Variability. Published doctoral dissertation, Universiti Teknologi Malaysia, Malaysia.

Santner, T.J., and Duffy, D.E. (1986). A note on A. Albert's and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 73(3):755-758.

Sarkar, S.K., Midi, H., and Rana, S.M. (2011). Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Sciences*. 11(1):26-35.

Satman, M.H., 2005. Outlier detection methods in linear regression. Social Sciences Institute, Istanbul University, Turkey

Schaefer, R.L. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*. 25(1-2):75-91.

Schaefer, R.L. (1979). Multicollinearity in logistic regression. Published doctoral dissertation, University of Michigan, USA.

Schaefer, R.L., Roi, L.D., and Wolfe, R.A. (1984). A ridge logistic estimator. *Communications in Statistics-Theory and Methods*. 13(1):99-113.

Segerstedt, B., and Nyquist, H. (1992). On the conditioning problem in generalized linear models. *Journal of Applied Statistics*. 19(4):513-526.

Shahmandi, M., Farmanesh, F., Gharahbeigi, M.M., and Shahmandi, L. (2013). Data analyzing by attention to weighted multicollinearity in logistic regression applicable in industrial data. *British Journal of Applied Science and Technology*. 3(4):748-763.

Shao, J. (1992). Bootstrap variance estimators with truncation. *Statistics and Probability Letters*, 15(2) : 95-101.

Shen, J., and Gao, S. (2008). A solution to separation and multicollinearity in multiple logistic regression. *Journal of Data Science*. 6(4):515-531.

Silvapulle, M.J. (1981). On the existence of maximum likelihood estimates for the binomial response models. *Journal of the Royal Statistical Society B*. 43(3), 310-3.

Šimecková, M. (2005). Maximum weighted likelihood estimator in logistic regression. Proceedings of Contributed Papers: Part 1 (pp. 144-148). Prague:MatfyzPress

Simpson, J.R. (1995). New Methods and Comparative Evaluations for Robust and Biased-Robust Regression Estimation. Published doctoral dissertation, Arizona State University, USA.

Smidt, R.K., and McDonald, L.L. (1976). Ridge Discriminant Analysis . Technical Report (No. S-1976-549, pp.1-32). Wyoming Univ Laramie Statictics Lab, Fort Belvoir, VA.

Stefanski, L.A., Carroll, R.J., and Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika*. 73(2):413-424.

Stromberg, A.J., and Ruppert, D. (1992). Breakdown in nonlinear regression. *Journal of the American Statistical Association*. 87(420):991-997.

Stromberg, A.J. (1997) Robust covariance estimates based on resampling. *Journal of Statistical Planning and Inference*, 57(2): 321-334.

Thisted, R.A. (1988). *Elements of statistical computing: numerical computation*. London:Chapman and Hall.

Tiku M.L., Tan, W.Y., and Bala-krishnan, N. (1986) *Robust Inference*. New York:Marcel Dekker.

Tukey, J.W. (1975). Mathematics and the picturing of data. Proceedings of the International Congress of Mathematics, vol. 2, pp. 523–531, August 1975.

Urgan, N.N., and Tez, M. (2008). Liu estimator in logistic regression when the data are collinear. Proceedings of the International Conference, EURO mini conference: Continuous Optimization and Knowledge-Based Technologies. (pp. 323-327).

Vago, E., and Kemeny, S. (2006). Logistic ridge regression for clinical data analysis (a case study). *Applied ecology and environmental research*, 4(2):171-179.

Vandev, D.L., and Neykov, N.M. (1998). About regression estimators with high breakdown point. *Journal of Theoretical and Applied Statistics*. 32(2): 111-129.

Vellman, P.F., and Welsch, R.E. (1981). Efficient computing of regression diagnostics. *Journal of American Statistics*. 35(4):234-242.

Victoria-Feser, M.P. (2002). Robust inference with binary data. *Psychometrika*. 67(1):21-32.

Walker, E. (1989). Detection of collinearity-influential observations. *Communications in Statistics-Theory and Methods*. 18(5):1675-1690.

Wang, C.Y., and Carroll, R.J. (1993). On robust estimation in logistic case-control studies. *Biometrika*. 80(1): 237-241.

Wang, C.Y., and Carroll, R.J. (1995). On robust logistic case-control studies with response-dependent weights. *Journal of Statistical Planning and Inference*. 43(3):331-340.

Wedderburn, R.W.M. (1976). On the existence and uniqueness of the maximum likelihood estimates forcertain generalized linear models. *Biometrika*. 63(1):27-32.

Weissfeld, L.A., and Sereika, S.M. (1991). A multicollinearity diagnostic for generalized linear model. *Communications in Statistics-Theory and Methods*. 20(4):1183-1198.

Willems, *G.,* Aelst, S. V. (2005) Fast and robust bootstrap for LTS,. 48(4):703-715

Williams, D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Journal of Applied Statistics*. 36(2):181-191.

Wu, J., and Asar, Y. (2016). On almost unbiased ridge logistic estimator for the logistic regression model. *Journal of Mathematics and Statistics*. 45(3):989-998.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*. 13(2):157-170.

BIODATA OF STUDENT

Syaiba Balqish binti Ariffin @ Mat Zin was born on July 5, 1978, in Kota Bharu city, as the second daughter of late Haji Ariffin Ismail and Hajah Sjarkiah Muhd Yasie. She took her secondary education at Sekolah Menengah Dato' Ahmed Maher, Kota Bharu in 1991 (listed as one of the Cluster Schools of Excellence in Malaysia). She was not only academically inclined, but was also active in her favorite clubs; St. John Ambulance, Shito-Ryu Karate-Do and Mural Art. In 1995, she was offered to join fast-track matriculation studies in Universiti Putra Malaysia, Mengkabang Telipot, Terengganu. She received the B.Sc. (Hons.) degree in Statistics from the Universiti Putra Malaysia, Serdang, Selangor in 2001, respectively. She was a former lecturer at Institute Technology Darul Naim, Kota Bharu and involved as a Biostatistician in Universiti Sains Malaysia, health campus, Kubang Kerian from 2003 to 2007. In April 2010, she received the M.Sc. Degree in Statistics from the same university, with a thesis on Robust Diagnostics in Logistics Regression Model. Later in 2012, she enrolled as a Ph.D. candidate at Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, continuing her scientific interest in robust statistics and logistic regression model.

LIST OF PUBLICATIONS

Ariffin, S.B. and Midi, H. (2018). Weighted high leverage collinear robust ridge estimator in logistic regression model. Pakistan Journal of Statistics, 34(1), 55-75.

Midi, H., **Ariffin, S.B.,** and Imon, A.H.M.R. (2016). Robust bootstrap procedure for estimation of binary logistic regression model in the presence of high leverage points with medical applications. International Journal of Applied Mathematics and Statistics, 53(1), 10-32.

PRESENTATIONS IN
INTERNATIONAL CONFERENCES & PROCEEDINGS

Ariffin, S.B., and Midi, H. Multivariate high leverage points in binary response for defect classification. The International Quantitative Research and Application Conference (IQRAC), 5th -8th August 2018, Kuching Sarawak, Malaysia.

Ariffin, S.B., and Midi, H. The effect of high leverage collinearity influential observations on logistic regression analysis. The 8th International Conference on Numerical Optimization and Operations Research (ICNOOR8), 27th -29th August 2017, Bukit Tinggi Sumatera Barat, Indonesia.

Ariffin, S.B., and Midi, H. On the Performance of High Leverage Collinearity Influential Diagnostic in Logistic Regression Model. The 6th International Conference on Numerical Optimization and Operations Research (ICNOOR6), 2nd – 4th September 2015, Siem Reap, Cambodia.

Ariffin, S.B., Midi, H., Rana, M.S., and Arasan, J. Robust Logistic Regression Multivariate High Leverage Points in Computer Engineering. The 3rd International Conference on Computer Engineering and Mathematical Sciences (ICCEMS3), 4th - 5th December 2014, Langkawi, Malaysia.

Ariffin, S.B., Midi, H., Arasan, J., and Rana, M.S. The effect of high leverage points on the maximum estimated likelihood for separation in logistic regression. The 2nd International Statistical Conference with Application in Sciences and Engineering (ISM2), 12th - 14th August 2014, Kuantan, Malaysia. AIP 2015, 1643(1), 402-408. DOI: 10.1063/1.4907472.

Ariffin, S.B., and Midi, H. Robust logistic ridge regression estimator in the presence of high leverage multicollinear observations. The 16th International Conference on Mathematical and Computational Methods in Science and Engineering (MACMESE16), 23rd – 25th April 2014, Kuala Lumpur, Malaysia.Pp. 179-184. WSEAS.

Ariffin, S.B., and Midi, H. The effect of high leverage points on the logistic ridge regression estimator having multicollinearity. The 3rd International Conference on Mathematical Sciences (ICMS3), 17th – 19th December 2013, Kuala Lumpur, Malaysia. AIP 2014, 1602(1), 1105-1111. DOI:10.1063/1.4882622

PRESENTATIONS IN NATIONAL CONFERENCES & PROCEEDINGS

Ariffin, S.B., Midi, H., and Arasan, J. Robust logistic ridge for multicollinearity in the presence of high leverage points. The 12th National Seminar of the Malaysian Statistical Institute (SKISM12), 29th March 2018, Universiti Putra Malaysia Serdang, Selangor.

Ariffin, S.B., Midi, H., Arasan, J. The effect of high leverage points on collinearity diagnostic in logistic regression model. The 9th Fundamental Science Congress Symposium of Mathematics (FSC9), 21st - 22nd November 2017, Universiti Putra Malaysia Serdang, Selangor.

Ariffin, S.B., Midi, H., Arasan, J., and Rana, M.S. Detection of High Leverage Influential Observations in Logistic Regression. The 10th National Seminar of the Malaysian Statistical Institute (SKISM10), 18th February 2016, Universiti Putra Malaysia Serdang, Selangor.

Ariffin, S.B., and Midi, H. Identification of High Leverage Collinearity Influential Observations in Logistic Regression Model. The 9th National Seminar of the Malaysian Statistical Institute (SKISM9), 31st March 2015, Universiti Kebangsaan Malaysia Bangi, Selangor.

Ariffin, S.B., Midi, H., Arasan, J., and Rana, M.S. The effect of high leverage point and separated data in the hidden logistic regression model. The 6th Fundamental Science Congress Symposium of Mathematics (FSC6), 18th – 19th August 2014, Universiti Putra Malaysia Serdang, Selangor.

Ariffin, S.B., Midi, H., Arasan, J., and Rana, M.S. Performance of ridge regression estimator for logistic regression model with application on cancer remission multicollinear data. The 5th Fundamental Science Congress Symposium of Mathematics (FSC5), 20th – 21st August 2013, Universiti Putra Malaysia Serdang, Selangor.

COMPETITIONS (3 MINUTES THESIS)

Outliers: A Matter of Survival? Inter-department (2nd Runner-up) and University heat, 2015.

Weak Statistics Implicated in Scientific Irreproducibility. Inter-department (1st Runner-up) and University heat, 2014.

Weak Data or Weak Statistics? Inter-department (1st Runner-up) and University heat, 2013.

UNIVERSITI PUTRA MALAYSIA

STATUS CONFIRMATION FOR THESIS / PROJECT REPORT AND COPYRIGHT

ACADEMIC SESSION : _____

TITLE OF THESIS / PROJECT REPORT :

ROBUST DIAGNOSTIC AND ESTIMATION FOR BINARY LOGISTIC REGRESSION MODEL IN THE PRESENCE OF MULTICOLLINEARITY AND HIGH LEVERAGE POINTS

NAME OF STUDENT: SYAIBA BALQISH BINTI ARIFFIN @ MAT ZIN

I acknowledge that the copyright and other intellectual property in the thesis/project report belonged to Universiti Putra Malaysia and I agree to allow this thesis/project report to be placed at the library under the following terms:

1. This thesis/project report is the property of Universiti Putra Malaysia.
2. The library of Universiti Putra Malaysia has the right to make copies for educational purposes only.
3. The library of Universiti Putra Malaysia is allowed to make copies of this thesis for academic exchange.

I declare that this thesis is classified as :

*Please tick (√)

☐	**CONFIDENTIAL**	(Contain confidential information under Official Secret Act 1972).
☐	**RESTRICTED**	(Contains restricted information as specified by the organization/institution where research was done).
☐	**OPEN ACCESS**	I agree that my thesis/project report to be published as hard copy or online open access.

This thesis is submitted for :

☐ **PATENT** Embargo from _____ (date) until _____ (date)

Approved by:

_____ (Signature of Student)
New IC No/ Passport No.:

Date :

_____ (Signature of Chairman of Supervisory Committee)
Name:

Date :

[Note : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization/institution with period and reasons for confidentially or restricted.]