2017-01-01

Analysis Of Bias-Corrected And Exact Estimators For Binomial Generalized Linear Model Parameters

Hamna Hannan
University of Texas at El Paso, hamnaikram17@gmail.com

ANALYSIS OF BIAS-CORRECTED AND EXACT ESTIMATORS FOR BINOMIAL GENERALIZED LINEAR MODEL PARAMETERS

HAMNA IKRAM

Master's Program in Statistics

APPROVED:

Amy E. Wagler, Ph.D., Chair

Dogan-Dunlap, Hamide, Ph.D.

Adeeba Raheem, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

to my

Dearest
MOTHER, FATHER
BROTHERS &
DAUGHTER

with love

ANALYSIS OF BIAS-CORRECTED AND EXACT ESTIMATORS FOR BINOMIAL GENERALIZED LINEAR MODEL PARAMETERS

by

HAMNA IKRAM

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Master's Program in Statistics

THE UNIVERSITY OF TEXAS AT EL PASO

August 2017

Acknowledgements

In the name of ALLAH, the most merciful, the most Gracious. All praise is due to ALLAH; we praise Him, seek His help and ask for His forgiveness. I am thankful to ALLAH Almighty, who supplied me with the courage, guidance and the love to complete this research.

As is often the case, this thesis would not have been possible without the help of a number of wonderful people whom I would like to take this opportunity to recognize. Foremost, of course, is my supervisor and mentor Dr. Amy Wagler, for investing her time, patience, guidance and friendship. I am lucky to have her as my teacher and thesis advisor. You are like a mother to me and have always supported and motivated me during these two years of my master's program. I am more grateful to her than she will ever know.

I wish to express special thanks to my committee members, Dr. Dogan-Dunlap, Hamide, Mathematical Sciences Department and Dr. Adeeba Raheem, clinical Assistant Professor in the Department of Civil Engineering, both at The University of Texas at El Paso. Their suggestions and comments were invaluable to the completion of this work.

My sincere thanks also go to Dr. Joan Staniswalis, Dr. Panagis Moschopoulos, Dr. Naijun Sha and Dr. Ori Rosen from the Mathematical Sciences Department at The University of Texas at El Paso who taught and mentored me. Additionally, I would like to thank all the professors, staff members, my friends (in El Paso and Pakistan), classmates and all my other well-wishers whose names are not mentioned for all their support. Last but not the least, I cannot forget Waleeja for her entire support.

I wish I had the words to express my deepest gratitude to my parents, my brothers; Aitsam, Ali and Zain and especially my loving daughter, Navera Fatima for her patience and sacrifice during my studies. Their prayers, love and constant support have made this journey possible for me and I dedicate my work to my parents.

Abstract

Typically, small samples have always been a problem for binomial generalized linear models. Though generalized linear models are widely popular in public health, social sciences etc. In small sample scenarios the non-existence of the maximum likelihood (ML) estimators is very common as well as separation occurs in the data. In logistic regression the maximum likelihood estimates are found to have biased away from origin. My work examines the bias-reduced and exact estimators that have been used to estimate the slope parameters and standard errors of the estimated slope parameters as compared to the traditional ML method.

The present work is noted for the logistic regression. For the models having categorical responses, bias-reduction performs the best. The latest research and methodological interest in the bias-reduction technique stimulate this current work and the main goal is to evaluate and broaden the application of this approach so it would identify the areas where bias-reduction can be helpful.

This research is an effort to prove theoretically and practically that bias-reduced method should be considered as an improvement over the traditional ML method. This method not only removes the first order bias in the ML method but also equivalent to the penalization of the likelihood by Jeffreys prior. Moreover, bias-reduced estimates are always find to be finite.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Regression analysis is one of the essential and the most widely used statistical method. Regression methods are used to investigate the relationship between a response (outcome) variable and one or more explanatory variables. Logistic regression models play an important role among different regression models. Under certain conditions [11], the linear regression model is a valuable tool for evaluating the effects of several explanatory variables on one dependent(outcome) continuous variable.

However, for situations where the dependent variable is qualitative and certain assumptions are not met, other methods have been developed. One of these is the logistic regression model, which deals with the case where the response is binary (discrete i-e non-normal), e-g. male vs. female, alive vs. dead, defective vs. non-defective etc. The details of the development on logistic regression in different areas have been provided by Agresti[2]. The term general linear model usually refers to the standard linear regression models in which response y is always continuous for given continuous/categorical predictors $x's$. It includes all the multiple regression models, ANOVA and ANCOVA (with fixed effects only). These models are based on the assumptions that all $y_i \sim N(iid)$, the distribution for all observations have constant variance i-e $\sigma_i^2 = \sigma^2 \ \ \forall i$ and the errors are normally distributed. Least squares and weighted least squares methods are used to fit these models. The general format of regression model is expressed with the $\boldsymbol{X\beta}$ term, where $\mathbf{X}$ is a known design matrix and ϵ is a random error.

$$\boldsymbol{Y_{n\times 1}} = \boldsymbol{X_{n\times p}}\ \boldsymbol{\beta_{p\times 1}} + \epsilon_{n\times p}$$

The term generalized linear models (GLM) extend the usual regression framework popularized by McCullagh and Nelder [14]. Generalized linear models are developed by relaxing

the assumptions of the standard linear regression model as mentioned above [11]. And these days GLMs are widely used in physical, biomedical, and behavioral sciences because of the non-normal responses. Example of GLM is the binary generalized model. For some particular scenarios the logistic regression coefficients estimates can be infinite i-e very large and the resultant estimates for parameters provided by the standard computer software are strange. As explained in Hosmer and Lemeshow (2004) [11] that not only the zero cells but also small samples cause separation problems for the usual maximum likelihood (ML) estimation. Also, Mehta and Patel(1995) [15] have shown the case by using real data when ML did not exist and he suggested an alternative approach "exact conditional" method. Moreover, we get very large standard errors when the parameter estimators are large. The above mentioned approaches are discussed in detail in chapter 2.

1.1 Separation

In such cases when the outcome variable in any data set is binary (discrete) and the number of data points is small, the situation of complete or quasi-separation easily arises within the observations. This happens when one or more of the independent variables can perfectly predict the dependent variable. Let's consider an example of diagnosis of breast cancer in females and males. The prediction of having the breast cancer or not is based on the gender. In such scenario there are more chances for females having the breast cancer than males. Thus, this example exhibits the case of complete separation. The maximum likelihood estimates will not converge either it is a case of complete separation or quasi separation. To illustrate the difference between complete and quasi-separation consider the following examples;

1.1.1 Complete Separation

When a linear function of the independent variable X_i can perfectly predict the response variable Y_i, a complete separation occurs. Perfect prediction implies that model cannot be

fitted and no parameter estimates can be obtained. As discussed in [3] Complete separation arises when there exists a (p+1) vector of coefficients $\boldsymbol{\beta}$ such that the outcome is $y_i = 0$ when $\boldsymbol{x_i\beta} \leq 0$ and when the outcome is $y_i = 1$ we have $\boldsymbol{x_i\beta} > 0$. The observations in

Table 1.1: Example of Complete Separation

i	1	2	3	4	5	6	7	8
x_i	-4	-3	-2	-1	1	2	3	5
y_i	0	0	0	0	1	1	1	1

Table 1.2: 2×2 contingency table for Complete Separation

X	Y	
	0	1
0	4	0
1	0	4

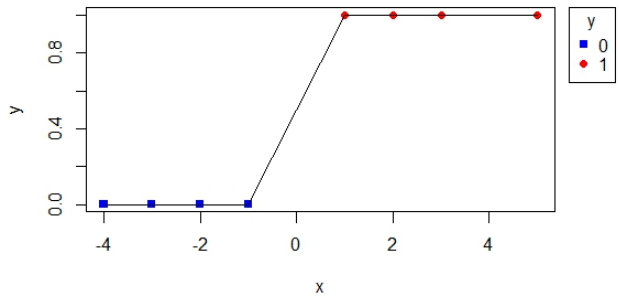


Figure 1.1: Scatter Plot under Complete Separation

the Table 1.1 represent a completely separated model having a single covariate X and the outcome variable Y. When $x_i \leq 0$, Y takes on the value $y_i = 0$ and whenever $x_i > 0$ then $y_i = 1$ for i = 1,2,...,8. The Information in Table 1.1 can be summarized in a 2×2

contingency table. The contingency Table 1.2 displays that off-diagonal cells have zero frequencies. That indicates complete separation in the data. Scatter plot is also helpful to represent complete separation between X and Y. By looking at the pattern of Fig. 1.1 we can see a sudden jump from $y_i = 0$ to $y_i = 1$

1.1.2 Quasi-complete Separation

Quasi-complete Separation in logistic regression arises when the outcome variable separates an explanatory variable or a combination of explanatory variables almost completely. Following Allison (2008) [3], Quasi-complete separation occurs when there exists a (p+1) vector of coefficients $\boldsymbol{\beta}$ such that $y_i = 0$ when $\boldsymbol{x_i\beta} \leq 0$ and $y_i = 1$ whenever $\boldsymbol{x_i\beta} \geq 0$ then and the equality holds for at least one category of the outcome variable. Table 1.3 represents the quasi-complete separation, where the single explanatory variable X and the outcome variable Y takes the value of $y_i = 0$ when $x_i < 0$ and $y_i = 1$ when $x_i > 0$. But, we can see a case where $y_i = 0$ and $y_i = 1$ has been observed for one value of X i-e $x_i = 0$, The observations in the Table 1.3 can be summarized by a 2×2 contingency table.

Table 1.3: Example of Quasi-complete Separation

i	1	2	3	4	5	6	7	8	9	10
x_i	-4	-3	-2	-1	0	0	1	2	3	5
y_i	0	0	0	0	0	1	1	1	1	1

Table 1.4: 2×2 contingency table for quasi Separation

X	Y	
	0	1
0	5	1
1	0	4

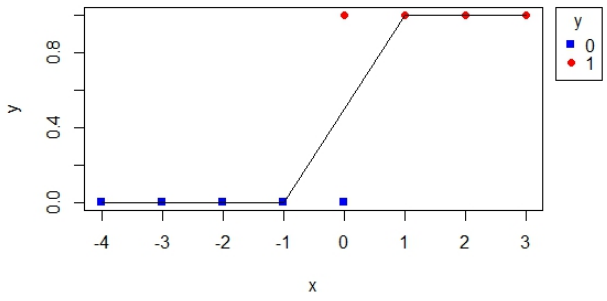


Figure 1.2: Scatter Plot under Quasi-Complete Separation

1.2 Bias

The bias of an estimator is defined as;

$$\boldsymbol{B}(\beta) = E_{\beta}(\hat{\beta} - \beta)$$

In a regular model with a p-dimensional parameter β the asymptotic bias [9] of the maximum likelihood estimate $\hat{\beta}$ is;

$$b(\beta) = \frac{b_1(\beta)}{n} + \frac{b_2(\beta)}{n^2} + ...$$

where "n" is the number of observations.

In statistics bias in estimation is always a concern for data analysists and researchers because the magnitude of bias plays very important role in estimation. If it is large the consequences can be misleading inferences. The aim of this research is to remove the $\boldsymbol{O}(n^{-1})$ term. For example, in the case of quantal logistic dose-response model Wagler(2011) [17] studied the bias in maximum likelihood (ML) estimator. The maximum likelihood (ML) estimator has asymptotically desirable behavior for the large samples and when the sample size is small the maximum likelihood (ML) estimators fail to provide the finite estimation of parameters, this is due to the bias in the maximum likelihood (ML) estimator. However, for small or moderate sample sizes the first-order bias term of ML estimator could be large. A lot of research has already been done on the methods that can be used for reducing the

amount of bias of the ML estimators and this present work is also a positive contribution towards the research. In order to reduce the bias present in the maximum likelihood estimates and the problem of non-convergence mainly two classes of methods are referred as “Bias correction” and “Bias reduction”.

1.2.1 Bias reducing methods

Firth(1993) [9] developed a method to remove the first order bias in the asymptotic expansion of the bias of the ML estimators (Detail of Firth’s method will be provided in chapter 2). In some important cases such as binomial logistic regression models, Poisson log-linear models and Gamma reciprocal-linear models the application of the method in generalized linear models (GLMs) with the canonical link is studied in Firth (1992a,b) [8], [7]. The bias corrected methods are based on the ML estimators and these estimators cannot be defined when the ML estimators are infinite. Many of them proposed [16] a very similar bias correction method based on the Lasso estimators. It has been proved that this method is equivalent in the second order to the bias reduction. But the bias reduction method does not depend upon the ML estimators. And the derived new estimators are found to have smaller first order term in the asymptotic expansion of their bias. In fact, this method is bias preventive rather than bias corrective.

1.3 Outline of the thesis

This current thesis is organized in the following manner. Chapter 2 provides a review of the literature on generalized linear models, exponential family and several aspects of the bias-reducing technique by modification of the score function. in chapter 3 we will use simulations to analyze and compare the performance of two methods other than usual maximum likelihood method for small samples. Furthermore, chapter 4 is the analysis of the methods by using a real life example. Finally, we will present and discuss the results obtained from chapters 3 and 4 and some recommendations.

Chapter 2

An Outline of Generalized Linear Models (GLMs)

Since the main focus of this research is the binomial generalized linear model but we will present some features of the generalized linear model first.

The generalized linear models are an extension of classical linear models as discussed in [14]. The GLM consists of three components i-e *Random Component*, *Systematic Component* and *Link function*. Generalized linear model can be expressed as;

$$E(Y_i/X_i) = g^{-1}(x_i^{'}\beta + \epsilon_i),\ i = 1, 2, ..., n \tag{2.1}$$

or,

$$\eta_i = g[E(Y_i/X_i)] = x_i^{'}\beta,\ i = 1, 2, ..., n \tag{2.2}$$

where g(.) is the link function that connects the expected response, $E(Y_i/X_i)$ to η_i, with x_i and β the p$\times$1 vectors of regression parameters, corresponding to Y_i and ϵ is identically, distributed random variable.

2.0.1 Likelihood function for generalized linear model

Each random component of Y_i in the generalized linear model has a distribution in the exponential family of the form, as discussed in (Agresti, P.133) [2];

$$f(y; \theta_i, \phi) = exp[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi). \tag{2.3}$$

The value of the parameter θ may vary for i = 1,2,... depending on values of explanatory variables. The parameter θ_i is called the natural parameter and ϕ is known as the dispersion parameter. If the distribution of the Y_i involves only unknown parameter as in binomial model then can take the value 1. When we model the count data for analysis in the form of proportion using a Bernoulli distribution for the probability of "success" and "failure" of a single trial, then for a fixed number of trials the binomial distribution will be suitable. A logistic function was developed in which the log odds of a success are expressed as linear combination of the covariates.

2.1 Binomial Logit Models for Binary Data

In this research, we are dealing with the binomial generalized linear model that fulfills the conditions of exponential-family because it can be written in the exponential-family form. The **Binomial** pdf is;

$$P(y_i) = \binom{n}{y} \pi^{y_i} (1-\pi)^{n-y_i}. \tag{2.4}$$

We represent the success and failure outcomes by 1 and 0 when the response variable is binary. The probabilities for success and failure are represented by $P(Y = 1) = \pi$ and $P(Y = 1) = 1 - \pi$, which is the special case of equation (2.4), when n = 1. The probability mass function for $y = 0$ *and* 1 is;

$$\begin{aligned} f(y;\pi) &= \pi^y (1-\pi)^{1-y} \\ &= (1-\pi)\left[\frac{\pi}{1-\pi}\right]^y \\ f(y;\pi) &= (1-\pi) exp\left[ylog\frac{\pi}{1-\pi}\right]. \end{aligned} \tag{2.5}$$

The equation (2.5) is the natural exponential family, where θ is identified with π, $a(\pi) = 1 - \pi$, $b(y) = 1$, and $Q(\pi) = log[\frac{\pi}{1-\pi}]$. The natural parameter $log[\frac{\pi}{1-\pi}]$ is the log odds of response 1, the *logit* of π, known as the canonical link and the GLMs using the logit link are called logit models.

2.1.1 Relationship between the probability and the odds of an event

The ratio between the number of successes and failures is called the odds of an event.

$$odds = [\frac{\pi_i}{1-\pi_i}] = \frac{P(Y_i = 1)}{P(Y_i = 0)} \tag{2.6}$$

and π_i is the probability of success, $0 \leq \pi_i \leq 1$ and the odds of an event will be greater than or equal to 0. The following table illustrates the the relationship between the odds of an event and the probabilities π_i. The Table 2.1 illustrates that when $0 < \pi_i < 0.5$ provides

Table 2.1: Relationship between probability, odds and log(odds)

$\boldsymbol{\pi_i}$	odds $= [\frac{\pi_i}{1-\pi_i}]$	$log[\frac{\pi_i}{1-\pi_i}]$
0	0	Doesn't exist
0.25	0.3333	-0.4771
0.5	1	0
0.75	3	0.4771
1	Not defined	Doesn't exist

odds $<$ 1and the log(odds) < 0, the value $\pi_i = 0.5$ gives odds $= 1$ and log(odds) $= 0$ and finally, $0.5 < \pi_i < 1$ yields odds > 1 and log(odds) > 0. The logit model (Cox and Snell, page 26) [6], can be written in the form of row vector x_i and column vector β as;

$$log[\frac{\pi_i}{1-\pi_i}] = \eta_i = x_i\beta \tag{2.7}$$

By taking the antilog of equation (2.7) get,

$$\begin{aligned} \frac{\pi_i}{1-\pi_i} &= e^{x_i\beta} \\ \pi_i &= \frac{e^{x_i\beta}}{1+e^{x_i\beta}} \end{aligned} \tag{2.8}$$

and,

$$1 - \pi_i = \frac{1}{1 + e^{x_i\beta}}. \tag{2.9}$$

Equation (2.9) can also be written as;

$$\pi_i = P(Y_i = 1) = \frac{1}{1 + e^{-x_i\beta}}$$

The figures, Fig. 2.1 and Fig. 2.2 (see Agresti, p.123) [2], depict the logistic curve of the probability π_i at different values of x_{ij} when β_j is positive and β_j is negative. Fig. 2.1 shows an upward trend when $\beta_j > 0$ and there is a downward trend in the Fig. 2.2 when $\beta_j < 0$.

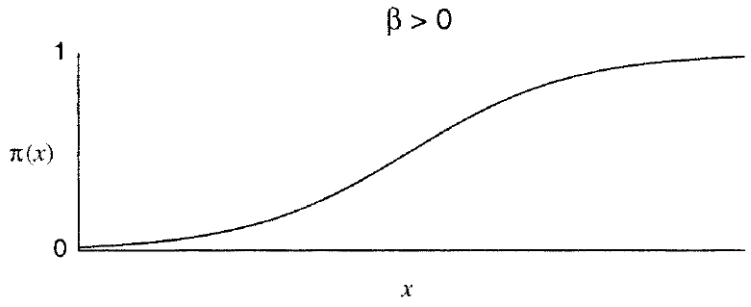


Figure 2.1: Logistic Regression Function

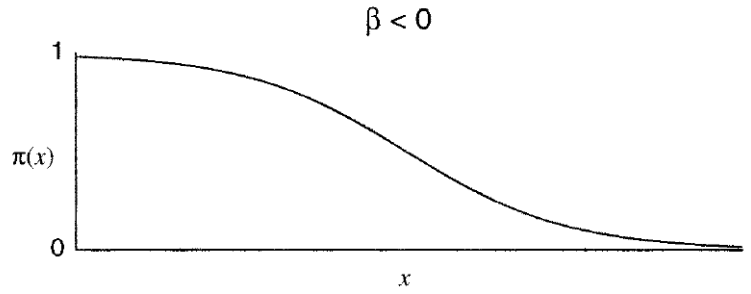


Figure 2.2: Logistic Regression Function

2.2 Methods to estimate parameters

We will use three methods i-e *Maximum Likelihood Estimation, Firth's Method and Exact Conditional Logistic Regression* but maximum likelihood (ML) very popular and commonly used method of estimation in the frequentist school. The easy implementation of fitting procedures and asymptotic properties of the ML estimator (unbiasedness, efficiency, asymptotic sufficiency and asymptotic normality) has made this method popular. Following (Agresti, P.195)[2] we let y_i represent the success count, then $(y_1, ..., y_n)^T$ are independent binomials.

$$= \prod_{n=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{n_i - y_i}. \tag{2.10}$$

Using equation (2.10) we can write the likelihood function for the n binary observations as a function of β.

$$\begin{aligned} L(\beta) &= P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n / \pi_1, \pi_2, ..., \pi_n) \\ &= \prod_{n=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1 - y_i}. \end{aligned} \tag{2.11}$$

Rewriting the equation (2.7) to obtain the log-likelihood for β as, $log(\pi_i) = x_i\beta + log(1-\pi_i)$ i-e

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \{y_i log\pi_i + (1 - y_i) log(1 - \pi_i)\} \\ &= \sum_{i=1}^{n} \{y_i(x_i\beta + log(1 - \pi_i)) + (1 - y_i) log(1 - \pi_i)\} \\ &= \sum_{i=1}^{n} \{y_i x_i \beta + log(1 - \pi_i)\} \\ &= \sum_{i=1}^{n} \{y_i x_i \beta - log(1 + exp(x_i\beta)\}. \end{aligned} \tag{2.12}$$

By taking the derivative of equation (2.7) and setting this derivative equal to 0 the ML estimate of β_j; $j = 0, 1, ..., p$ can be obtained.

$$U(\beta_j) = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{exp(x_{ij}\beta_j)(x_{ij})}{1 + exp(x_{ij}\beta_j)} \tag{2.13}$$

From equation (2.12) substituting the value of π_i as,

$$\pi_i = \frac{exp(x_{ij}\beta_j)}{1 + exp(x_{ij}\beta_j)}$$

Gives the value of $U(\beta_j)$ as,

$$\begin{aligned} U(\beta_j) &= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \pi_i x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \pi_i) x_{ij} = 0 \ j = 0, 1, ..., p. \end{aligned} \tag{2.14}$$

There is no closed form of the solution for the equation (2.14) therefore, requires iterative solution.

2.2.1 Newton Raphson Method

Following [4] let x_i be the i^{th} row vector $x_i = (1, x_{i1}, ..., x_{ip})$ as (p+1)×1 column vector of the covariates for observations i, i-e x^T. The first derivative of the log-likelihood w.r.t β can be found as;

$$\begin{aligned} U(\beta) &= \frac{\partial log L(\beta)}{\partial \beta} \\ &= \sum_{i=1}^{n} x y_i - \sum_{i=1}^{n} x(1 + e^{x_i \beta})^{-1} \end{aligned} \tag{2.15}$$

In equation (2.15) $U(\beta)$ is called the *score* function or sometimes known as *gradient*.

$$\begin{aligned} I(\beta) &= \frac{\partial^2 log L(\beta)}{\partial \beta \partial \beta^T}. \\ &= -\sum_{i=1}^{n} x_i x_i^T (1 + e^{\beta x_i^T})^{-1} (1 - (1 + e^{x_i^T \beta})^{-1}). \end{aligned}$$

where, $I(\beta)$ is the matrix of second derivative and is also called the Hessian matrix. The Hessian matrix is also used to find the variance and covariance of the estimated regression coefficients. As illustrated in [11] the diagonal and off-diagonal elements of the the Hessian matrix can be written as;

$$\frac{\partial^2 l(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i).$$

and,

$$\frac{\partial^2 l(\beta)}{\partial\beta_j\partial\beta_l} = -\sum_{i=1}^{n} x_{ij}x_{il}\pi_i(1-\pi_i).$$

To solve the problem of separation and to reduce the bias that is found in the ML estimates because of the small samples, we will use two methods, i-e Firth's Method (Bias Reduction Method) and Conditional Logistic Regression. But, first we will discuss the Firth's Method

2.2.2 Firth's Method

Before Firth, Cordeiro and McCullagh (1991) [5] has presented that how the asymptotic bias in the ML estimator of order $O(n^{-1})$ could be eliminated by using the supplementary weighted regression. But they derived it in the case of the univariate generalized linear models for the asymptotic bias of the ML estimator. However, the bias-correction in the above-mentioned situation depends on the existence of ML estimates and it does not apply to the situation where the ML estimates are found to have infinite values. To a first order of approximation for large sample size it is shown that the ML estimates are unbiased in [14] that;

$$E(\hat{\beta}-\beta) = O(n^{-1}).$$

and asymptotic variance is the inverse of the Fisher matrix expressed as;

$$Cov(\hat{\beta}) = (X^T W(\beta)X)^{-1}\{1+O(n^{-1}\}.$$

Motivated partly, Firth (1993) [9] developed a method for the removal of the first order $O(n^{-1})$ term in the expansion of the bias of ML estimator discussed in chapter 1. This method is a modification of the score function (first derivative of the maximum likelihood). The Firth's method is different than the procedure mentioned above to remove the $\frac{b_1(\beta)}{n}$ term from the asymptotic bias. Because in the bias corrected method the parameter is first calculated then it is corrected but in Firth's method, the corrective procedure is applied to the ML estimate before calculating the parameter estimate. As explained in (Agresti,

P.134) [2] the exponential form of the binomial distribution is;

$$f(y_i;\pi_i,n_i) = \binom{n_i}{n_i y_i}\pi_i^{n_i y_i}(1-\pi_i)^{n_i - n_i y_i}$$
$$= exp\left[\frac{y_i\theta_i - log[1+exp(\theta_i)]}{1/n_i} + log\binom{n_i}{n_i y_i}\right]. \tag{2.16}$$

The equation (2.16) takes the form of exponential family with $b(\theta_i) = log[1+exp(\theta_i)]$, $a(\phi) = 1/n_i$, and $c(y_i,\phi) = log\binom{n_i}{n_i y_i}$. Logit, $\theta_i = log[\pi_i/1-\pi_i]$ is the natural parameter.

2.2.3 Modified Score Function

The ML estimate [9] for the parameter β_j is derived by taking the derivative of the log likelihood $l(\beta)$, *score function* and set it equal to 0,

$$U(\beta_j) = \frac{\partial log l(\beta)}{\partial \beta_j} = 0 \tag{2.17}$$

By the definition of the sufficient statistic t_j and the log likelihood function the coefficient β_j can be expressed as;

$$l(\beta_j) = \sum_{i=1}^{n}\left\{y_i x_{ij}\beta_j - log(1+e^{x_{ij}\beta_j})\right\}.$$

where, $l(\beta_j) = logL(\beta_j)$

$$= t_j\beta_j - K(\beta_j).$$

The score function in equation (2.14) can be expressed as;

$$U(\beta_j) = \tfrac{\partial logL(\beta)}{\partial \beta_j}.$$
$$= \sum_{i=1}^{n}(y_i - \hat{\pi}_i x_{ij})x_{ij}.$$
$$= t_j - K^{'}(\beta_j). \tag{2.18}$$

which explains that the sufficient statistic t affects only the location of $U(\beta)$, not its curvature. As, explained in [9] it can be shown that bias in β_j appears from the combination

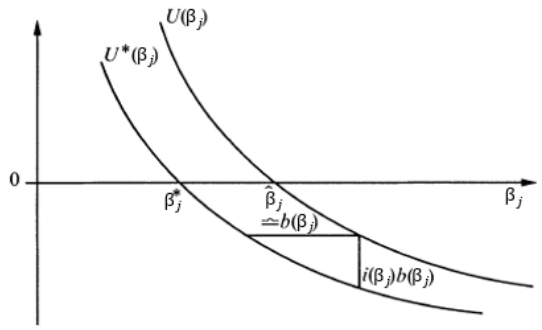


Figure 2.3: Modified Score Function

of unbiasedness of the score function, $E\{U(\beta)\} = 0$ at the true value of β_j and due to the curve in the score function, as depicted from the Figure, $U^{''}(\beta_j) = K^{''}(\beta_j) \neq 0$. The idea of this work is to illustrate that the bias in $\hat{\beta}$ can be reduced by adding a small bias into the score function. From the basic triangle geometry, represented in Fig. 2.3 the score function is shifted downward by amount of $i(\beta_j)b(\beta_j)$ if the estimator $\hat{\beta}_j$ has positive bias of $b(\beta_j)$, where the gradient is expressed as $U^{'} = -i(\beta_j)$ and modified score function of the modified estimate β_j^* can be obtained by by setting $U^*(\beta_j) = 0$.

$$U^*(\beta_j) = U(\beta_j) - i(\beta_j)b(\beta_j)$$

2.2.4 Jeffreys Invariant Prior and Modified Score Function

By Jeffreys invariant prior [12] the likelihood function in equation (2.11) can be penalized to get the modified score function. Zorn [18] also suggested that separation problem can be dealt by maximizing the penalized likelihood rather than the usual likelihood function. The Jeffreys prior is $|I(\beta|^{1/2} = |X^T W(\beta)X|^{1/2}$ where $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ is the vector of unknown parameters and the Fisher information matrix is $I(\beta) = X^T W(\beta)X$. Here,X is the design matrix and $W(\beta_j)$ is a n×n diagonal matrix written as; $W = diag[\pi_i(1 - \pi_i)]$.

The penalized log likelihood and likelihood function in [9] are described as;

$$l^*(\beta) = l(\beta) + \frac{1}{2} log|I(\beta)|.$$
$$L^*(\beta) = L(\beta) \times |I(\beta)|^{1/2}. \tag{2.19}$$

In equation (2.19) the penalty $|I(\beta)|^{1/2}$ is the Jeffreys [12] invariant prior. It has been explained in section (2.2.3) that θ is the canonical parameter of our binomial exponential family model and it has been discussed in section 2, (Firth, 1993) [9] that the $O(n^{-1})$ bias is removed by calculating the posterior mode based on the Jeffreys prior for the canonical parameter of an exponential family. Following [9] we can write;

$$U^*(\beta_j) = U(\beta_j) + \left(\frac{1}{2}\right) trace[(X^T W(\beta_j) X)^{-1} (X^T W(\beta_j) X)].; \ for \ j = 0, 1, ...p \tag{2.20}$$

The details of above proof are provided in appendix A.

2.3 Exact Conditional Logistic Regression

Another method in which the exact logistic regression estimates can be obtained even in the case of complete separation and empty cells is the Exact conditional logistic regression. This method is based on the same idea as for exact inference in 2×2 contingency tables. The methodology of exact conditional logistic regression was first suggested by (Cox and Snell, 1970) [6]. Following (King and Ryan, 2002) [13] exact conditional logistic regression inference is based on the exact permutational distributions of the sufficient statistics that correspond to the parameters of interest, conditional on fixing the sufficient statistics of the remaining parameters at their observed values [18]. However, the above-mentioned approach has some problems that will be discussed in the later chapters.

2.3.1 The ML estimate

In order to estimate the parameter we will use the unconditional likelihood function to find the conditional likelihood function. From equation (2.11) for binary observation, we have

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}. \\ &= \prod_{i=1}^{n} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i). \\ &= \prod_{i=1}^{n} e^{x_i y_i \beta}(1+e^{\eta_i})^{-1}, \quad \text{where, } \eta_i = x_i\beta \\ &= \frac{e^{\sum_{i=1}^{n}\sum_{j=0}^{p}\beta_j x_{ij} y_i}}{\prod_{i=1}^{n}(1+e^{\eta_i})}. \end{aligned} \tag{2.21}$$

Equation (2.21) can be written as;

$$L(\beta) = \frac{e^{\sum_{j=0}^{p}\beta_j t_j}}{\prod_{i=1}^{n}(1+e^{\eta_i})}. \tag{2.22}$$

Following (Hosmer and Lemeshow, P.332) [11] the form of likelihood function in equation (2.22) suggests that sufficient statistic for β_j is

$$t_j = \sum_{i=1}^{n} x_{ij} y_i.$$

for $j = 0, 1, ..., p$ and t_j is the sum of all explanatory variables for $y_i = 1$, and let $t^{'} = (t_1, ...t_p)$ is the vector of sufficient statistic for the slope coefficients. The exact distribution of the collection of p sufficient statistic is given by;

$$P(T_1 = t_1, T_2 = t_2, ..., T_p = t_p) = \frac{c(\mathbf{t})exp(\sum_{j=1}^{p} t_j\beta_j)}{\sum_{u\in S} c(\mathbf{u})exp(\sum_{l=1}^{p} u_l\beta_l)}. \tag{2.23}$$

In equation (2.23) the number of possible allocations of 0 and 1 to $(y_1, ..., y_n)$ are c($\mathbf{t}$), such that $t_j = \sum_{i=1}^{n} x_{ij}y_i$ and S represents the set of allocations of 0 and 1 to $(y_1, ..., y_n)$ such that $n_1 = \sum_{i=1}^{n} y_i$ and the resulting value of the jth statistic for the lth allocation are indicated by $u_j = \sum_{l=1}^{n} x_{ij}y_i$.

Chapter 3

Simulations

3.1 Background

In this chapter, I will evaluate the biasedness and finite sample performance of maximum likelihood estimators, biased-reduced estimators and exact conditional estimators of the logistic regression model via Monte Carlo simulations. This simulation study comprises of one covariate (predictor) x, which is fixed and the binary response y, is binomially randomly generated random variable. The general form of vector of parameters is $\beta = (\beta_0, \beta_1)^T$ but my focus is on slope parameter. I conducted simulations for three different values of the population parameter β, the covariate, x, and the sample size, n where $n = length(x)$. In the simulations, I considered three different values of $\beta = (0.1, 1, 3)$, sample sizes $n = 25, 50$ and 1000 independent sets of random samples for each sample size are generated. For each set of parameter and sample size, I estimated the slope of the regression parameter and standard errors of the slope parameter. The final estimates of β and $SE(\beta)$ are the average of the 1000 estimates of β correspond to that particular sample size. The simulated results using *MLE (ML), Bias-reduced (BR)* and *Exact conditional (EC)* methods are presented in the following table.

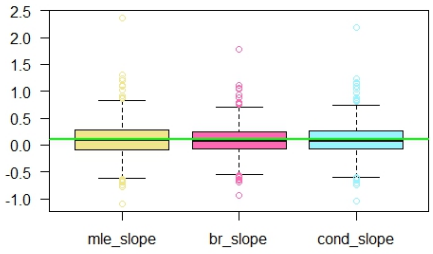
In the Table 3.1 the first row displays the binary regression estimates and standard error of slope estimates obtained by the maximum likelihood estimation. The second rows explains about the estimates obtained from bias reduction method and the third row provides the estimates by using the exact conditional method at different values of slope population parameters. The estimated values in the above table has been obtained after removing the outliers for different scenarios of β_i.

Table 3.1: Simulated parameter estimates and standard errors by using three methods

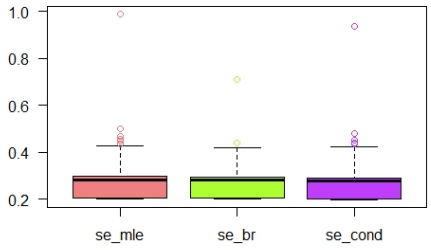
Methods	Parameter (True value)					
	β=0.1		β=1		β=3	
	Estimate	SE	Estimate	SE	Estimate	SE
MLE	0.11322	0.25523	1.14327	0.40177	2.63951	0.89643
Bias Reduced	0.10355	0.25277	1.03015	0.37250	2.67516	0.97273
Exact conditional	0.10954	0.25085	1.09635	0.38899	2.49510	0.85051

In the above table for the parameter $\beta = 1$ we see that the bias reduction method almost provides the zero bias though the slope estimates using two other methods are also close to the true value but keep this thing in mind that at this point 26 outliers have been removed from the ML and exact conditional method. More explanation will be provided with the boxplots. For the second case when $\beta = 0.1$ all the slope estimates are very close to each other and no outlying values were detected. For the last scenario when $\beta = 3$ the ML and exact conditional model did not converge properly and this is because of the separation problem for the small data set. There were 958 cases that were detected as outliers for MLE and Exact conditional methods. It is interesting to note that for the large value of β the bias from all three methods are under estimated. Since, slope estimates are not indicative of separability in the data so our main focus in this research is the standard errors of the estimated regression coefficients.

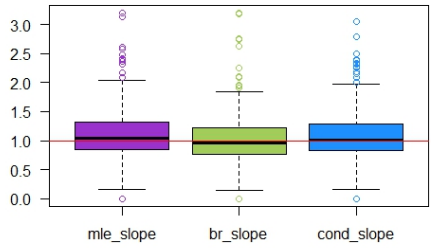
The boxplots (a & b) in the Fig. 3.1 display the slope estimates and their standard errors side by side. When $\beta = 0.1$ we could not find very large standard errors and this is due to the fact that slope parameter is small. This 0.1 value of β does not provide large log odds because of the weak association. For such a small value of β, it does not matter a lot which approach should be used but bias reduced method is the best to pick. Let's consider the case for $\beta = 1$. We explored some cases where the slope estimates exceeded the value 2. This is the only rule of thumb to set a cut-off for slope estimates a value 2. It is important to note that the large estimated values of slope also an indication of the inflated



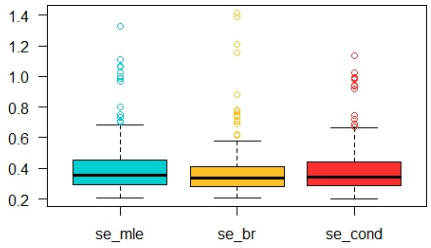
(a) Boxplot of $\hat{\beta}_{0.1}$



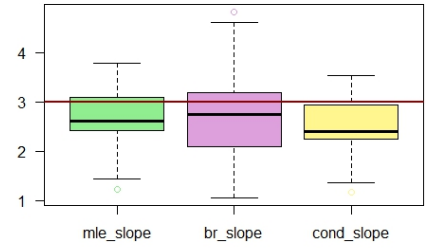
(b) Boxplot of SE($\hat{\beta}_{0.1}$)



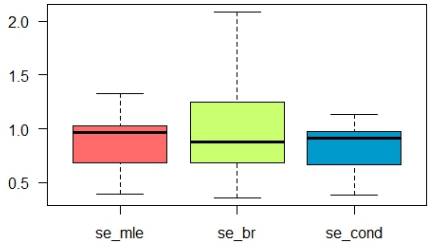
(c) Boxplot of $\hat{\beta}_1$



(d) Boxplot of SE($\hat{\beta}_1$)



(e) Boxplot of $\hat{\beta}_3$



(f) Boxplot of SE($\hat{\beta}_3$)

Figure 3.1: Comparison of estimated slopes and standard error of estimated regression coefficients

standard errors and some of the cases were noted where the standard errors exceeded 6000. There were 26 cases of both methods i-e ML and Exact conditional that we removed. It is worthwhile, to note that the value of $\beta = 1$ also provides large log odds as compared to $\beta = 0.1$. The situation for $\beta = 3$ is not different than the $\beta = 1$ and there were 958 points identified as outlying points. It is quite interesting that ML and exact conditional provided the same results and BR resulted in the small standard errors. I used a cut-off hundred for standard errors as a rule of thumb but in chapter 5 we have derived a justified lower bound on mean square errors (MSE) of $\hat{\beta}$.

3.2 Summary

Based upon the above-simulated results and discussion we can analyze that there separation in the small samples also related to the value of the slope parameter. The larger the value of the parameter the stronger will be the association. Let's consider the scenario of strong association for the highest value of $\beta = 3$ which accounts for the quasi-separation in the data. The summary of the three methods in Table 3.1 helps us in deciding that even if we exclude some outlying points that the ML and exact conditional estimators take infinite values, the BR estimator has overall smallest bias and standard error. Therefore, the bias reduction (Firth) method is the best amongst all. The bias reduction method corrects up to second order, whereas in the MLE there is no correction available and this is the reason we see those cases which are inestimable.

Chapter 4

Application

For demonstrative purposes, we illustrate the application of the different methodologies discussed in chapter 3, by using a real life data example. We want to see what would happen if we were introduce two continuous predictors here in the example. The endometrial cancer data set has been taken from Heinze and Schemper(2002) [10] and link of the data set is provided in the appendix C.

4.1 Data Analysis

The endometrial data see [10] & [1] involves predicting the histology on the basis of three risk factors or predictors (covariates). The study consists of 79 histology cases i-e binary response y_i, classified as either "0" (low grade for 30 patients) or "1" (high grade for 49 patients) respectively with 3 covariates, Nevasculation (NV), Pulsatility index (PI) and Endometrium height (EH) which are the supposed three risk factors. PI and EH are continuous covariates. Three risk factors were considered; x_1 = neovascularization (NV) is coded as 1 (present) for 13 patients and 0 (absent) for 66 patients, and two continuous factors, x_2 = pulsatility index of arteria uterina (PI) and x_3 = endometrium height (EH) range from 0 to 49 and from 0.27 to 3.61. We will use the following model to analyze our endometrial data set.

$$logit[p(Y = 1)] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

The data has been analyzed using R software (for R-codes see appendix C.) and summary of the three approaches used for analysis is displayed in the Table 4.1. In the Table, 4.1 the output for the usual MLE and bias reduced method is in the log of odds ratio while exact

Table 4.1: Endometrial Data Analysis

	MLE	Bias-reduced	Exact conditional
$\hat{\beta}_{NV}$	18.1856	2.9293	$1.144e^{+09}$
SE($\hat{\beta}_{NV}$)	1715.7508	1.55076	$1.106e^{+04}$
$p-value_{NV}$	0.9915	0.0589	$0.9985e^{0}$
$\hat{\beta}_{PI}$	-0.0422	-0.0348	$9.595e^{-02}$
SE($\hat{\beta}_{PI}$)	0.0443	0.03958	$4.388e^{-02}$
$p-value_{PI}$	0.3413	0.3799	$0.3457e^{0}$
$\hat{\beta}_{EH}$	-2.9026	-2.6042	$5.815e^{-02}$
SE($\hat{\beta}_{EH}$)	0.8456	0.7760	$8.333e^{-01}$
$p-value_{EH}$	0.00060	0.00080	$0.00064e^{0}$

conditional illustrates the odd ratios of the estimated slope coefficients and their standard errors.

By looking at the Table 4.1 we can interpret that for usual ML, the NV is causing quasi-separation in the data because of the extremely inflated value of SE($\hat{\beta}_{NV}$) that leads to an insignificant p-value of NV. Although, the estimated $\hat{\beta}_{NV}$ is showing kind of relationship but still this value is large. As stated above that neovascularization is an important risk factor, so it is not sensible decision to drop it from the model because of its infinite estimate. Hence, the ML method cannot be used in this scenario. We attempted the exact conditional here but this does not necessarily converge to a solution. Moreover, the output showed some warnings for the nonconvergence issues.

The bias reduced (Firth's method) version clearly shows the improvement on the estimates as well as their standard errors also reduced as compared to the other two methods applied before. It provides the best results in order to reduce the bias that arises in the small samples and in the separated data. All the estimated coefficients and their standard errors are finite and it proves that bias reduced method eliminates the bias up to order $O(n^{-2})$ which traditional ML method cannot do.

As explained in the previous chapters that problems of separation occur in the small samples or when there is an extreme split on the response variable. We can conclude on the

basis of the above data analysis output that our proposed method *Bias Reduced* is very helpful when quasi-complete separation occurred in the space of explanatory variables.

Chapter 5

Discussion and Conclusion

The work done by [16] on the bias adjustment to minimize the asymptotic mean square error for an ML estimator motivated me to find some numerical bounds on the bias adjustment that reduce the high-order mean square error for a logit model with small samples. As explained in [16] this bias adjustment provides the mean square error that is smaller than that we get from the the maximum likelihood and asymptotically unbiased estimator.

The inflation in standard errors were seen in the simulation results for different values of the population slope parameters and as a consequence the solutions were non-converged for those cases. The literature provided in [16] intuited me to investigate that what numerical value would expect to see in a well converging solution. For details see appendix C. When we are working with the log odd ratios then subtracting off the bias-corrections makes the analysis complicated so using this idea is advantageous to multiply the estimator by the penalty "c^*". The MSE of $c^*\hat{\beta}^*$ is;

$$\begin{aligned} MSE(c^*\hat{\beta}^*) &= E[c^*\hat{\beta}^* - \beta_0]^2 \\ &= E(c^{*2}\hat{\beta}^{*2}) - 2c^*\hat{\beta}^*\beta_0 + {\beta_0}^2 \\ &= c^{*2}[var(\hat{\beta}^* + {\beta_0}^2] - 2c^*{\beta_0}^2 + {\beta_0}^2 \end{aligned} \tag{5.1}$$

In order to get the c^*_{min}, we take the partial derivative of equation 5.1 and set it equal 0.

$$\frac{\partial MSE(c^*\hat{\beta}^*)}{\partial c^*} = 2c^*[var(\hat{\beta}^* + {\beta_0}^2] - 2{\beta_0}^2 + 0$$

$$c^*[var(\hat{\beta}^* + {\beta_0}^2] = {\beta_0}^2$$

$$c^*_{min} = \frac{{\beta_0}^2}{var(\hat{\beta}^* + {\beta_0}^2)}; < 1$$

$$c^*_{min} = \left\{1 + {\beta_0}^{-2} var(\hat{\beta}^*)\right\}^{-1} \tag{5.2}$$

$$MSE(c^*\hat{\beta}^*) = n^{-1}\beta_2 + n^{-2}(\beta^*_{\Delta 2} - {\beta_0}^{-2}{\beta_2}^2) + O(n^{-3}) \tag{5.3}$$

Let's consider a case by taking $n = 50$, $\beta_0 = -1.099$, $\beta_2 = 5$, $\beta^*_{\Delta 2} = \left[\frac{1}{-(\pi)^2} + \frac{1}{(1-\pi)^2}\right]$, where $\pi = 0.25$.

where; β_0 is population slope parameter, β_2 is the second order asymptotic variance, $\beta^*_{\Delta 2}$ second derivative of the higher order asymptotic variance and in our case $c^* = 1$ and the left hand side in our equation is the mean square error of bias reduced estimator. By plugging the values in equation 5, we get a value 0.1004. This helps us in understanding that for larger values of the asymptotic variance this bound increases but at a much lower rate. We can also see the rise in the higher order asymptotic variance and a decreasing effect too that helps in minimizing the inflation in the higher order variance.

5.1 Significance of the Result

The main purpose of the work done in this thesis how the magnitude of bias can be reduced under separated data in cases dealing with binary outcomes for small samples applying different methods used come with some restrictions. This thesis has led to the development a lower bound on the MSE of parameter estimates for varying population parameters.

From the simulation results using single covariate and the practical example consisting of categorical or continuous covariates bias reduced method performed the best in comparison with the usual ML and exact conditional method. Furthermore, the bias reduced estimator

always take finite values even in the circumstances when there is complete separability in the data.

5.2 Recommendations

To an experimenter who wants to get the best under situations of small samples and separation when analyzing the binary responses, we propose the following;

1. It is not necessary that not all the logistic regression models will converge. Sometimes, the software results can be misleading. See table 4.1. It is advised to check the finiteness of estimator and the standard errors of that estimator.

2. The large values of slope parameters give rise to big odd ratios. Bias Reduced is the best approach to use because it shrinks the estimator towards the true parameter value by reducing the amount of bias.

3. In the visual analysis of the real world data it is recommended to look at the pattern at the plot and if it is making the pattern then quasi-separation is detected in the data set.

Bibliography

[1] A. Agresti. *Categorical Data Analysis.* Wiley Series in Probability and Statistics. Wiley, 2013.

[2] Alan Agresti. *Categorical data analysis*, volume 990. New York: John Wiley & Sons, 1996.

[3] Paul D Allison. Convergence failures in logistic regression. In *SAS Global Forum*, volume 360, pages 1–11, 2008.

[4] P.D. Allison. *Logistic Regression Using SAS: Theory and Application, Second Edition.* ITPro collection. SAS Institute, 2012.

[5] Gauss M. Cordeiro and Peter McCullagh. Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):629–643, 1991.

[6] D.R. Cox and E.J. Snell. *Analysis of Binary Data, Second Edition.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.

[7] D Firth. Generalized linear models and jeffreys priors: an iterative weighted least-squares approach. In *Computational statistics*, pages 553–557. Springer, 1992.

[8] David Firth. Bias reduction, the jeffreys prior and gum. *Advances in GLIM and Statistical Modelling*, 13:91, 1992.

[9] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[10] Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419, 2002.

[11] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Applied Logistic Regression. Wiley, 2004.

[12] Harold Jeffreys. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London a: mathematical, physical and engineering sciences*, volume 186, pages 453–461. The Royal Society, 1946.

[13] Elizabeth N King and Thomas P Ryan. A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician*, 56(3):163–170, 2002.

[14] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.

[15] Cyrus R Mehta and Nitin R Patel. Exact logistic regression: theory and examples. *Statistics in medicine*, 14(19):2143–2160, 1995.

[16] Haruhiko Ogasawara. Bias adjustment minimizing the asymptotic mean square error. *Communications in Statistics-Theory and Methods*, 44(16):3503–3522, 2015.

[17] A Wagler. Bias reduction in logistic dose-response models. *Journal of biopharmaceutical statistics*, 21(3):405–422, 2011.

[18] Christopher Zorn. A solution to separation in binary response models. *Political Analysis*, 13(2):157–170, 2005.

Appendix A

The n×n matrix $W(\beta)$ in chapter 2 is defined as;

$$W(\beta) = \begin{bmatrix} \pi_1(1-\pi_1) & 0 & 0 & \dots & 0 \\ 0 & \pi_2(1-\pi_2) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \pi_n(1-\pi_n) \end{bmatrix}$$

From equation (2.20);

$$W(\beta_j) = \text{diag}\left\{ \frac{\partial}{\partial \beta_j} \{(\pi_i(1-\pi_i))\} \right\}$$

where, $\pi_i = (1+e^{x_{ij}\beta_j})^{-1}$

$$W(\beta_j) = \text{diag}\left\{ \frac{\partial}{\partial \beta_j} \{(1+e^{x_{ij}\beta_j})^{-1}(1+e^{-x_{ij}\beta_j})^{-1}\} \right\}$$

By taking partial derivative w.r.t β_j we get,

$$\begin{aligned} W(\beta_j) &= \text{diag}\left[x_{ij} \left\{ \frac{-(1+e^{x_{ij}\beta_j}) + (1+e^{-x_{ij}\beta_j})}{(1+e^{x_{ij}\beta_j})^2 + (1+e^{-x_{ij}\beta_j})^2} \right\} \right] \\ &= \text{diag}\left[x_{ij} \left\{ \frac{-1}{(1+e^{x_{ij}\beta_j}) + (1+e^{-x_{ij}\beta_j})^2} + \frac{1}{(1+e^{x_{ij}\beta_j})^2 + (1+e^{-x_{ij}\beta_j})} \right\} \right] \\ &= \text{diag}\left[x_{ij} \left\{ \frac{-1}{(1+e^{x_{ij}\beta_j}) + (1+e^{-x_{ij}\beta_j})^2} + \frac{1}{(1+e^{x_{ij}\beta_j})^2 + (1+e^{-x_{ij}\beta_j})} \right\} \right] \\ &= \text{diag}\left(x_{ij}(\pi_i)(1-\pi_i)(-\pi_i + 1 - \pi_i) \right) \\ &= \text{diag}\left(x_{ij}(\pi_i)(1-\pi_i)(1-2\pi_i) \right) \end{aligned}$$

Now, the equation (2.20) can be written as;

$$U^*(\beta_j) = U(\beta_j) + \left(\frac{1}{2}\right) trace[W(\beta)X(X^T W(\beta_j)X)^{-1}X^T Z(\beta_j)].$$

Which is equivalent to the 'hat' matrix in [9];

$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}.$$

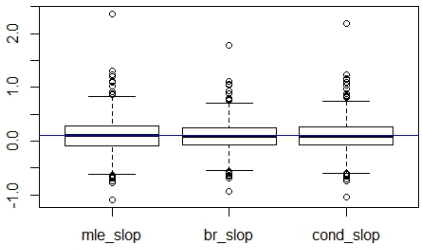
Now, the equation (A) becomes;

$$\begin{aligned} U^*(\beta_j) &= U(\beta_j) + (\frac{1}{2})\sum_{i=1}^{n}(1-2\pi_i)h_i x_{ij} \\ &= \sum_{i=1}^{n}(y_i - \pi_i)x_{ij} + (\frac{1}{2})\sum_{i=1}^{n}(1-2\pi_i)h_i x_{ij} \\ &= \sum_{i=1}^{n}\left\{y_i + (h_i/2) - \pi_i - h_i\pi_i\right\}x_{ij} \end{aligned}$$

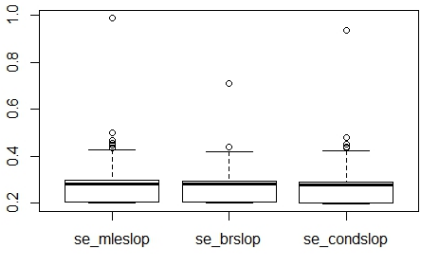
Appendix B

```
reps = 1000
set.seed(100)
results_new = matrix(NA, reps*6, 16)

index=0

require(brglm)
require(survival)

for(b in c(1, 0.1,3))
{
  for (r in c(5,10))
  {
    for (i in 1: reps)
    {

      x = rep(c(-2,-1,0,1,2), each = r)
      eta = x * b
      p = exp(eta)/(1+exp(eta))
      n = length(x)
      y= rbinom(n,size=1,prob=p)
      reg = data.frame(y,x)
      MLE = tryCatch(glm(y ~ x ,data = reg,  family = "binomial"))
      br = tryCatch(brglm( y ~ x, data = reg,  family = "binomial"))
```

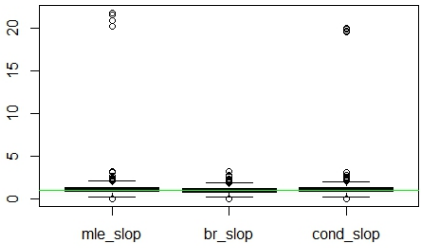
```
      cond = try(clogit(y ~ x, data= reg))

      se = summary(MLE)$coefficients[,2]
      (beta0MLE_se = se[1])
      (beta1MLE_se = se[2])
      se = summary(br)$coefficients[,2]
      (beta0br_se = se[1])
      (beta1br_se = se[2])
      se = summary(cond)$coefficients[,3]
      (beta1cond_se = se[1])

      index = index+1
      results_new[index,] =   c(b,r, MLE$coefficients[1], MLE$coefficients[2],
       AIC(MLE), br$coefficients[1],br$coefficients[2],
      br$coefficients[2], AIC(br), cond$coefficients[1] ,
      beta0MLE_se, beta1MLE_se,beta0br_se, beta1br_se,
        beta1cond_se,AIC(cond),n)colnames(results_new, do.NULL = FALSE)
       colnames(results_new) = c("par","r", "mle_b0","mle_slop"
       ,"aic_mle","br_b0","br_slop","aic_br","cond_slop","se_mleb0",
       "se_mleslop","se_pmleb0",
       "se_brslop","se_condslop","aic_cond","n");results_new
}}}


###  MLE
mle1_slope = new_beta1[ ,1]
mean(mle1_slope)
summary(mle1_slope)
```



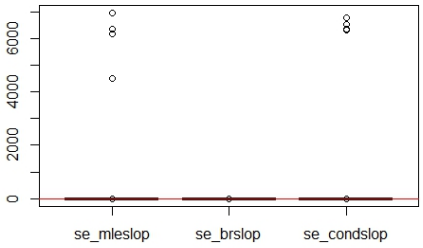
(a) Boxplot of $\hat{\beta}_{0.1}$



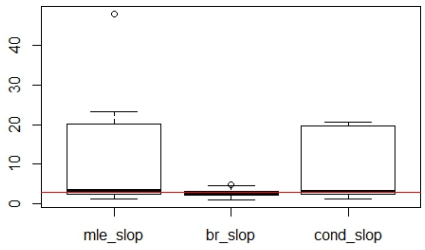
(b) Boxplot of SE($\hat{\beta}_{0.1}$)



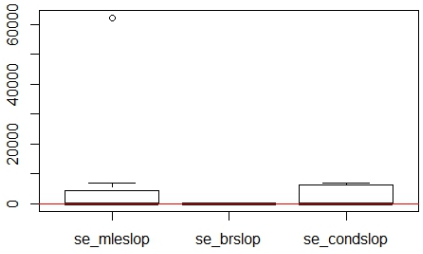
(c) Boxplot of $\hat{\beta}_1$



(d) Boxplot of SE($\hat{\beta}_1$)



(e) Boxplot of $\hat{\beta}_3$



(f) Boxplot of SE($\hat{\beta}_3$)

Figure B.1: Comparison of estimated slopes and standard error of estimated regression coefficients with outliers

```
mle10_slope = mle1_slope[mle1_slope < 10]
mean(mle10_slope)
summary(mle10_slope)

### Br

br1_slope=new_beta1[,2]
mean(br1_slope)
summary(br1_slope)


### conditional

cond1_slope = new_beta1[,3]
mean(cond1_slope)
summary(cond1_slope)

cond10_slope = cond1_slope[cond1_slope < 10]
mean(cond10_slope)
summary(cond10_slope)


boxplot(mle10_slope,br1_slope,cond10_slope,las=1,names=c("mle_slope",
"br_slope","cond_slope"),
col = c("darkorchid3","darkolivegreen3","dodgerblue1"),
outcol=c("darkorchid3",
"darkolivegreen3","dodgerblue1"))
abline(h =1, col = "red", lwd = 1)
```

```
###  MLE

mle0.1_slope= new_beta0.1[,1]
mean(mle0.1_slope)
summary(mle0.1_slope)

### Br

br0.1_slope = new_beta0.1[,2]
mean(br0.1_slope)
summary(br0.1_slope)
### conditional

cond0.1_slope =new_beta0.1[,3]
mean(cond0.1_slope)
summary(cond0.1_slope)



boxplot(mle0.1_slope,br0.1_slope,cond0.1_slope,las=1,names=c("mle_slope"
,"br_slope","cond_slope"),
col = c("khaki", "hotpink", "cadetblue1"),outcol=c("khaki",
"hotpink", "cadetblue1"))
abline(h =0.1, col = "green", lwd = 2)


###  MLE
```

```
mle3_slope=new_beta3[ ,1]
mean(mle3_slope)
summary(mle3_slope)

mlecut_slope = mle3_slope[mle3_slope < 10]
mean(mlecut_slope)
summary(mlecut_slope)

### br

br3_slope=new_beta3[ ,2]
mean(br3_slope)
summary(br3_slope)

### conditional

cond3_slope = new_beta3[ ,3]
mean(cond3_slope)
summary(cond3_slope)

condcut_slope = cond3_slope[cond3_slope < 10]
mean(condcut_slope)
summary(condcut_slope)

boxplot(mlecut_slope,br3_slope,condcut_slope,las=1,names=c("mle_slope",
"br_slope","cond_slope"),
  col=c("lightgreen", "plum", "khaki1"),outcol=c("lightgreen",
```

```
   "plum", "khaki1"))
abline(h =3, col = "darkred", lwd = 2)


#### Standard errors

###  MLE

se.out_mle1= se_beta1[ ,1]
mean(se.out_mle1)
summary(se.out_mle1)

cutm1_se = se.out_mle1[se.out_mle1 < 100]
mean(cutm1_se)
summary(cutm1_se)

### Br_SE

se.out_br1= se_beta1[ ,2]
mean(se.out_br1)
summary(se.out_br1)

### conditional

se.out_cond1= se_beta1[ ,3]
mean(se.out_cond1)
summary(se.out_cond1)
```

```
cutc1_se = se.out_cond1[se.out_cond1 < 100]
mean(cutc1_se)
summary(cutc1_se)

boxplot(cutm1_se,se.out_br1,cutc1_se,las=1,names=c("se_mle",
"se_br","se_cond"),
col=c("darkturquoise", "goldenrod1", "firebrick1"),
outcol=c("darkturquoise","goldenrod1", "firebrick1"))


###  MLE

se.out_mle0.1=se_beta0.1[ ,1]
mean(se.out_mle0.1)

### Br_se

se.out_br0.1 = se_beta0.1[ ,2]
mean(se.out_br0.1)
summary(se.out_br0.1)

### conditional

se.out_cond0.1 = se_beta0.1[ ,3]
mean(se.out_cond0.1)
summary(se.out_cond0.1)
```

```
boxplot(se.out_mle0.1,se.out_br0.1,se.out_cond0.1,las=1,
names=c("se_mle","se_br","se_cond"),
col=c("lightcoral", "greenyellow", "darkorchid1"),
outcol=c("lightcoral","greenyellow", "darkorchid1"))


###  MLE

se.out_mle3 = se_beta3[ ,1]
mean(se.out_mle3)
summary(se.out_mle3)


cutm3_se = se.out_mle3[se.out_mle3 < 100]
mean(cutm3_se)
summary(cutm3_se)

### Br

se.out_br3 = se_beta3[ ,2]
mean(se.out_br3)
summary(se.out_br3)

### conditional
cutc3_se = se.out_cond3[se.out_cond3 < 100]
mean(cutc3_se)
summary(cutc3_se)
```

```
boxplot(cutm3_se,se.out_br3,cutc3_se,las=1,names=c("se_mle","se_br","se_cond"),
col=c("indianred1", "darkolivegreen1", "deepskyblue3"),outcol=c("indianred1",
 "darkolivegreen1", "deepskyblue3"))
```

Appendix C

```
http://www.stat.ufl.edu/~aa/cda/data.html

require(brglm)
require(survival)

## Traditional MLE
endo.glm <- glm(HG ~ NV + PI + EH, data = endomet,family = binomial)
summary(endo.glm)

## brglm

endo.brglm <- brglm(HG ~ NV + PI + EH, data = endomet,family = binomial)
summary(endo.brglm)

## Exact Conditional Logistic Regression
endo.clogit <- clogit(HG ~ NV + PI + EH, data = endomet)
summary(endo.clogit)
```

Minimizing the MSE by multiplying the unbiased estimator by a constant c^*. Define $\hat{\beta}^*$ as an estimator with $\mathrm{E}(\hat{\beta}^*) = \beta_0$ corresponding to the biased $\hat{\beta}$, then the MSE of $c^*\hat{\beta}^*$ is;

$$\begin{aligned}
MSE(c^*\hat{\beta}^*) &= E[c^*\hat{\beta}^* - \beta_0]^2 \\
&= E[c^*\hat{\beta}^* - E(\hat{\beta}^*) + E(\hat{\beta}^*) - \beta_0]^2 \\
&= E[(c^*\hat{\beta}^* - \beta_0)^2 + (\beta_0 - \beta_0)^2 + 2(c^*\hat{\beta}^* - 0)(0)] \\
&= E\left(c^{*2}\hat{\beta}^{*2} + \beta_0^2 - 2c^*\hat{\beta}^*\beta_0\right) \\
&= c^{*2}E(\hat{\beta}^{*2}) + \beta_0^2 - 2c^*\beta_0 E(\hat{\beta}^*) \\
MSE(c^*\hat{\beta}^*) &= c^{*2}[Var(\hat{\beta}^*) + \beta_0^2] - 2c^*\beta_0^2 + \beta_0^2 \qquad \text{(C.1)}
\end{aligned}$$

By taking the partial derivative of equation C.1, we find c^*_{min};

$$\begin{aligned}
\frac{\partial}{\partial c^*} MSE(c^*\hat{\beta}^*) &= 2c^*[Var(\hat{\beta}^*) + \beta_0^2] - 2\beta_0^2 + 0 = 0 \\
c^*[Var(\hat{\beta}^*) + \beta_0^2] &= \beta_0^2 \\
c^*_{min} &= \frac{\beta_0^2}{Var(\hat{\beta}^*) + \beta_0^2} \\
c^*_{min} &= \left(1 + \beta_0^{-2} Var(\hat{\beta}^*)\right)^{-1}
\end{aligned}$$

We get;

$$\begin{aligned}
c^*_{min} &= 1 - \beta_0^{-2}Var(\hat{\beta}^*) + [\beta_0^{-2}Var(\hat{\beta}^*)]^2 + O(n^{-3}) \\
&= 1 - \beta_0^{-2}\frac{b_2}{n} + \beta_0^{-4}Var(\hat{\beta}^*)^2 + O(n^{-3}) \\
c^*_{min} &= 1 - n^{-1}\beta_0^{-2}\beta_2 + n^{-2}(\beta_0^{-4}\beta_2^2 - \beta_0^{-2}\beta^*_{\Delta 2}) + O(n^{-3}) \ < 1
\end{aligned}$$

The quantity $\beta^*_{\Delta 2}$ is the higher order asymptotic variance of $\hat{\beta}^*$ as explained in [16]. The MSE minimized among the class of $c^*\hat{\beta}^*$ is;

$$\begin{aligned}
MSE(c^*\hat{\beta}^*) &= \frac{Var(\hat{\beta}^*)}{1 + \beta_0^{-2}Var(\hat{\beta}^*)} \\
&= \frac{n^{-1}\beta_2 + n^{-2}\beta^*_{\Delta 2}) + O(n^{-3})}{1 + n^{-1}\beta_0^{-2} + O(n^{-2})} \\
&= n^{-1} + n^{-2}(\beta^*_{\Delta 2} - \beta_0^{-2}\beta_2^2) + O(n^{-3}) \ < Var(\beta^*).
\end{aligned}$$

Curriculum Vitae

Hamna Ikram was born on September 17, 1982, in Sahiwal, Punjab, Pakistan.After, completing senior high school, she continued her college education in Govt. Postgraduate Degree college for Women, Sahiwal where she earned her bachelor's degree in Mathematics, Statistics, and Economics with First Division. She was recognized as outstanding student for the Governor Scholarship.

Thereafter, she enrolled in Graduate School at the University of Texas at El Paso in Fall 2015 to pursue a Master's degree in Statistics. During the Master's degree, she served as a Teaching Assistant and a tutor at the Mathematics Resource Center for students (MaRCS). She plans to continue her studies in Ph.D. Statistics in the near future.

Contact Information: hhannan@miners.utep.edu