Penalized Logistic Regression in Case-Control Studies

by

Jiying Wen

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2016

Approval

Name:	**Jiying Wen**
Degree:	**Master of Science (Statistics)**
Title:	***Penalized Logistic Regression in Case-Control Studies***
Examining Committee:	**Chair:** Rachel Altman Associate Professor

Jinko Graham
Senior Supervisor
Professor _____

Tim Swartz
Supervisor
Professor _____

Brad McNeney
Internal Examiner
Associate Professor _____

Date Defended: 16 December 2016

Abstract

Likelihood-based inference of odds ratios in logistic regression models is problematic for small samples. For example, maximum-likelihood estimators may be seriously biased or even non-existent due to separation. Firth proposed a penalized likelihood approach which avoids these problems. However, his approach is based on a prospective sampling design and its application to case-control data has not yet been fully justified.

To address the shortcomings of standard likelihood-based inference, we describe: i) naive application of Firth logistic regression, which ignores the case-control sampling design, and ii) an extension of Firth's method to case-control data proposed by Zhang. We present a simulation study evaluating the empirical performance of the two approaches in small to moderate case-control samples. Our simulation results suggest that even though there is no formal justification for applying Firth logistic regression to case-control data, it performs as well as Zhang logistic regression which is justified for case-control data.

Keywords: Logistic regression; case-control data; small samples; separation; profile likelihood

Dedication

To my beloved parents, for their endless support and encouragement.

Acknowledgements

First and foremost, I want to express my sincerest gratitude for my senior supervisor, Dr. Jinko Graham. I would like to thank her for her generous support and guidance throughout my time at SFU, without which the project could not be completed so smoothly.

I would like to acknowledge the Department of Statistics and Actuarial Science at Simon Fraser University. My graduate experience benefited greatly from the courses I took, and the high-quality seminars that the department organized.

I would also take the opportunity to thank Dr. Brad McNeney, Dr. Tim Swartz and Dr. Rachel Altman for being my project committee, generously giving their time and expertise to be my project readers and participate in my defense.

I extend my gratitude to my fellow graduate students, especially the 2015 cohort, for the stimulating discussions, the sleepless nights we were working together before deadlines, and all the fun we have had in the last one and half years. Special thank to Trevor Thomson, with whom I have shared many hours, discussions, outings, much laughter, struggles or satisfaction of the graduate school life.

Above all else, I am forever grateful to my parents, who were always willing to lend an ear to my troubles, encourage me when everything seems hopeless, and cheer me on when I succeed. I could not have made it this far without their unconditional love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

A case-control study is an observational epidemiological study that is designed to help examine the association between a risk factor and disease. Generally, a case-control study compares subjects with the disease of interest (cases) to a suitable control group of subjects without the disease (controls), and looks back retrospectively to learn which subjects in each group had the exposure to the risk factor, comparing the frequency of the exposure in the case group to the control group.

The major difference between cohort and case-control study designs is in the selection of the study subjects. In a cohort study, we start by selecting a fixed number of subjects who are initially free of disease and classify them according to their exposure to the risk factors of interest, then follow up to see how many develop the disease. In a case-control study, we identify subjects on the basis of presence or absence of the disease of interest, then trace back to investigate their past exposure to putative risk factors.

The ratio of controls to cases is an important design decision to consider in conducting a case-control study. In unmatched case-control studies, the optimal control-to-case ratio would be roughly 1:1, i.e. one case to one control, if the number of available cases and controls is large and the cost of obtaining information from both groups is comparable (dos Santos Silva, 1999). However, sometimes the number of cases available for the study is small and cannot be increased - e.g., 8 cases of vaginal adenocarcinoma among young woman in Boston area in the 1971 case-control study of Herbst and colleagues. In this situation, an equally small number of controls would provide little ability to find associations. The control-to-case ratio can be increased to ensure that the study will have the necessary statistical power to be able to detect an effect. For a given number of cases, the more

controls per case, the greater the statistical power of the study. However, when the control-to-case ratio is beyond 4:1, the marginal increase in statistical power with each additional control is negligible (dos Santos Silva, 1999).

Herbst et al. (1971) used a 4:1 case-control study design to examine the effect of *in-utero* exposure of dietheylstibestrol (DES) on the subsequent development of vaginal adenocarcinoma, a very rare disease, among young women. The study included eight cases of vaginal cancer in women aging from 15 to 22 years, and each case was matched with four controls that were born within 5 days at the same hospital in the same type of room (either public or private) as the case. They compared the use of DES by their mothers during pregnancy to see if the treatment with DES was more common among mothers of the cases. Among the mothers of the eight cases, seven had received DES during pregnancy, while none of the mothers of the controls had taken DES (see Table 4.1 in the Application section).

For a rare disease such as vaginal adenocarcinoma, a case-control study is the only reasonable approach to identify the causative agent. The researchers could have performed a prospective cohort study to get a group of women whose mothers used DES while pregnant and another group of women whose mothers did not, and observed these groups for a period of time for the disease development. However, given how uncommon the disease outcome is, even a large prospective study would have been unlikely to have more than one or two cases, even after 15-20 years of follow-up. Therefore, such prospective studies are impractical for rare diseases, expensive, and perhaps even unethical. In contrast, a case-control study starts with people known to have the outcome of interest, rather than starting with a population free of disease and waiting to see who develops it. Therefore, case-control studies are particularly suitable for the study of risk factors related to low incidence (i.e. rare) diseases with long induction periods.

The standard method for analysis of case-control studies is logistic regression corresponding to the prospective model for the probability of disease given covariates, which ignores the retrospective design. The parameters are usually estimated using maximum likelihood estimation (MLE) via numerical methods such as the Newton-Raphson algorithm. The desirable properties of MLE such as consistency, efficiency and normality, are based on the assumption that the sample size approaches infinity. However, in many real life case-control studies, the sample size is relatively small and so the large sample assumption is not satisfied. As a result, the usual MLEs of the log odds ratio parameters in logistic regression are biased. Moreover, in small-sample, case-control studies of a rare disease, there is a non-negligible probability of encountering the problem of separation, in which a single covariate, or a linear combination of several risk factors, can "separate" the binary outcome (Albert and Anderson, 1984) as described in the next paragraph. This type of data has

been shown to have the property of monotone likelihood. In this situation, the likelihood function has no maximum and the MLE does not exist.

Separation occurs in the DES example described above. Let the binary covariate DES denote the prenatal DES exposure status of the subjects, with $DES = 1$ for exposed and $DES = 0$ for non-exposed. Let D be the binary outcome variable for disease status. Based on the data, we have $DES \leq 0$ (in fact $DES = 0$) for all controls and $DES \geq 0$ for all cases. In this sense, the DES covariate separates the cases and controls. As there is no observation for which $DES = 1$ and $D = 0$, so-called "quasicomplete" separation occurs, leading to an infinite ML estimate of the effect of DES.

One of the possible remedies for separation is a penalized likelihood approach known as Firth logistic regression. For prospective data, Firth (1993) proposed a general method to reduce the small-sample bias in the MLE by introducing a small bias in the score function, and noted that solving the modified score equations is equivalent to maximizing a penalized likelihood. His approach leads to an estimator for β that is equivalent to the Bayesian posterior mode under the Jeffreys prior distribution (Jeffreys, 1946). The Firth logistic regression estimator always exists, even under separation, and is unique. However, this penalized likelihood method was built on prospective rather than retrospective likelihoods. As the problem of separation is so common in small-sample case-control studies, it is very tempting to apply Firth logistic regression to case-control data. However, to the best of our knowledge, Firth logistic regression for case-control data has not yet been fully justified.

1.2 Overview of the Project

I begin with a basic overview of the logistic regression model and how it can be applied to the case-control data. I then describe the principle of Firth logistic regression, a penalized likelihood method that researchers might use to handle separation in the case-control studies. Next, I summarize a profile-likelihood approach that incorporates a Firth-like correction proposed by Zhang (2006) for case-control data. I use a simulation study to evaluate the empirical performance of corresponding estimators derived from the aforementioned two methods, followed by application of these approaches to the DES data introduced in Section 1.1. To end the project, I provide some concluding remarks.

Chapter 2

Methods

2.1 Logistic Regression and Case-Control Data

2.1.1 The Logistic Regression Model

Logistic regression is a commonly used tool to describe the relationship between a binary outcome variable and a set of explanatory variables. The popularity of logistic regression stems mainly from its mathematical convenience and the relative ease of interpretation in terms of odds ratios.

For the ith subject, assume that the outcome variable y_i is Bernoulli distributed and takes on the value 1 with probability $\pi_i = P(y_i = 1|\boldsymbol{x}_i)$, where $\boldsymbol{x_i} = (x_1, \ldots, x_p)$ is the subject's covariate vector, and value 0 with probability $1-\pi_i$. The logistic regression model can be written as:

$$\pi_i = \frac{\exp(\alpha + \boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + \exp(\alpha + \boldsymbol{x}_i^T\boldsymbol{\beta})}, \tag{2.1}$$

where α is an intercept term, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p\times 1$ vector of odds ratio parameters.

Equation (2.1) gives a generalized linear model. Let $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})^T$, then the likelihood function and the corresponding log-likelihood function are given by the following equations:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \text{and}$$
$$\ell(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\}.$$

One of the most popular methods to estimate the unknown coefficients $\boldsymbol{\theta}$ is maximum likelihood (ML) estimation. In order to find the value that maximizes $\log \mathcal{L}(\boldsymbol{\theta})$, partial

derivatives of the log-likelihood function with respect to $\boldsymbol{\theta}$ are calculated as follows:

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = U(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \pi_i)\boldsymbol{x}_i. \tag{2.2}$$

The second derivative with respect to $\boldsymbol{\theta}$ of the log likelihood function, or the Hessian matrix, can be expressed as

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -\sum_{i=1}^{n} \boldsymbol{x_i}^T \boldsymbol{x_i} \pi_i (1 - \pi_i).$$

The solution to the score equation $U(\boldsymbol{\theta}) = 0$ gives the ML estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$. There is no analytical solution to the score equations; therefore, numerical methods (e.g. Newton-Raphson or Fisher Scoring) are used to find $\hat{\boldsymbol{\theta}}$. With the starting value $\boldsymbol{\theta}^{(1)}$, $\hat{\boldsymbol{\theta}}$ is obtained iteratively until the convergence of parameter estimates. The iterative Newton-Raphson algorithm is defined as:

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} + \mathbf{I}^{-1}(\boldsymbol{\theta}^{(r)}) U(\boldsymbol{\theta}^{(r)}), \tag{2.3}$$

where the superscript (r) denotes the number of the iteration, and $\mathbf{I}(\boldsymbol{\theta})$ denotes the Fisher information matrix, i.e., the expected value of minus the second derivative of the log likelihood, evaluated at $\boldsymbol{\theta}$. In the context of logistic regression,

$$\mathbf{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \mathbf{X}^T \mathbf{W} \mathbf{X},$$

where $\mathbf{X}$ is an $n \times (p+1)$ design matrix with elements in the first column being 1, and $\mathbf{W}$ is an $n \times n$ diagonal matrix with general element $\pi_i(1 - \pi_i)$.

Asymptotically, the MLEs $\hat{\boldsymbol{\theta}}$ are normally distributed around the true parameter $\boldsymbol{\theta}$, and the estimated variance-covariance matrix, $\text{Var}(\hat{\boldsymbol{\theta}})$, is obtained by evaluating the inverse of the Fisher information matrix $\mathbf{I}^{-1}$ at the MLEs, with the standard errors of single parameters corresponding to the diagonal elements of the matrix.

2.1.2 Likelihood for Case-Control Data

Under a cohort design, in which we assume that the values of the covariates are fixed and the outcome is then measured conditionally on the observed values of the covariates, the logistic regression model is simply equation (2.1), where the exposure $\boldsymbol{x}$ is treated as a fixed quantity, and the response variable y is random.

In a case-control study, we have selected cases and controls from the population; therefore, the binary outcome variable is fixed by stratification, and the exposure variables are then measured for each subject selected. The following development of the likelihood func-

tion for case-control data is based on the arguments of Hosmer and Lemeshow (2000). As the likelihood function for case-control data is based on subjects selected, it is necessary to define a variable that records the selection status for each subject in the population. Let the variable s denote the selection ($s = 1$) or non-selection ($s = 0$) of a subject. Let $p_1 = P(s = 1|y = 1)$ denote the probability of sampling a case, $p_0 = P(s = 1|y = 0)$ denote the probability of sampling a control. For a sample of n_1 cases ($y = 1$) and n_0 controls ($y = 0$), by Bayes rule, we can rewrite the expression of p_1 and p_0 as

$$\begin{aligned} p_1 = P(s = 1|y = 1) &= \frac{P(y = 1|s = 1)P(s = 1)}{P(y = 1)} = \frac{n_1}{n_1 + n_0} \cdot \frac{P(s = 1)}{P(y = 1)}, \text{ and} \\ p_0 = P(s = 1|y = 0) &= \frac{n_0}{n_1 + n_0} \cdot \frac{P(s = 1)}{P(y = 0)} \end{aligned} \tag{2.4}$$

respectively.

The full likelihood for the sample of n_1 cases and n_0 controls is

$$\prod_{i=1}^{n_1} P(\boldsymbol{x_i}|y_i = 1, s_i = 1) \prod_{i=1}^{n_0} P(\boldsymbol{x_i}|y_i = 0, s_i = 1). \tag{2.5}$$

For an individual term in the likelihood function shown in equation (2.5), by Bayes theorem, we have

$$P(\boldsymbol{x}|y, s = 1) = \frac{P(y|\boldsymbol{x}, s = 1) \cdot P(\boldsymbol{x}|s = 1)}{P(y|s = 1)}. \tag{2.6}$$

By conditional probability, the first term in the numerator of equation (2.6), we get equation (2.7) when $y = 1$,

$$P(y = 1|\boldsymbol{x}, s = 1) = \frac{P(s = 1|\boldsymbol{x}, y = 1) \cdot P(y = 1|\boldsymbol{x})}{P(s = 1|\boldsymbol{x}, y = 1) \cdot P(y = 1|\boldsymbol{x}) + P(s = 1|\boldsymbol{x}, y = 0) \cdot P(y = 0|\boldsymbol{x})}. \tag{2.7}$$

Further assume that the selection of cases and controls is independent of the covariates with respective probabilities p_1 and p_0, we obtain

$$\begin{aligned} p_1 &= P(s = 1|y = 1, \boldsymbol{x}) = P(s = 1|y = 1), \text{ and} \\ p_0 &= P(s = 1|y = 0, \boldsymbol{x}) = P(s = 1|y = 0). \end{aligned}$$

Inserting p_0, p_1 and the standard logistic regression model, equation (2.1), into equation (2.7) produces

$$\begin{aligned}
P(y=1|\boldsymbol{x},s=1) &= \frac{p_1 \cdot P(y=1|\boldsymbol{x})}{p_1 \cdot P(y=1|\boldsymbol{x}) + p_0 \cdot P(y=0|\boldsymbol{x})} \\
&= \frac{p_1 \cdot \frac{\exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})}}{p_1 \cdot \frac{\exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})} + p_0 \cdot \frac{1}{1+\exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})}} \\
&= \frac{p_1 \cdot \exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})}{p_0 + p_1 \cdot \exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})} \\
&= \frac{\frac{p_1}{p_0} \cdot \exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})}{1 + \frac{p_1}{p_0} \cdot \exp(\alpha+\boldsymbol{x}^T\boldsymbol{\beta})} \\
&= \frac{\exp(\log(\frac{p_1}{p_0}) + \alpha + \boldsymbol{x}^T\boldsymbol{\beta})}{1 + \exp(\log(\frac{p_1}{p_0}) + \alpha + \boldsymbol{x}^T\boldsymbol{\beta})} \\
&= \frac{\exp(\alpha^* + \boldsymbol{x}^T\boldsymbol{\beta})}{1 + \exp(\alpha^* + \boldsymbol{x}^T\boldsymbol{\beta})},
\end{aligned}$$

where $\alpha^* = \alpha + \log\left(\frac{p_1}{p_0}\right)$. Substituting the expression of p_0 and p_1 defined in (2.4), we have

$$\alpha^* = \alpha + \log\left(\frac{n_1}{n_0}\right) - \log\left(\frac{\phi}{1-\phi}\right), \tag{2.8}$$

where $\phi = P(y=1)$ is the population probability of having the disease.

Further, let $\pi^*(\boldsymbol{x}) = P(y=1|\boldsymbol{x},s=1) = \frac{\exp(\alpha^*+\boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\alpha^*+\boldsymbol{x}^T\boldsymbol{\beta})}$. Assuming the sampling is carried out independently of covariate values, we have

$$\begin{aligned}
P(\boldsymbol{x}|y=1,s=1) &= \pi^*(\boldsymbol{x}) \cdot \frac{P(\boldsymbol{x}|s=1)}{P(y=1|s=1)} \\
&= \pi^*(\boldsymbol{x}) \cdot \frac{P(\boldsymbol{x})}{P(y=1|s=1)}.
\end{aligned}$$

By a similar argument, the corresponding result for $P(\boldsymbol{x}|y=0,s=1)$ is given by

$$P(\boldsymbol{x}|y=0,s=1) = [1-\pi^*(\boldsymbol{x})] \cdot \frac{P(\boldsymbol{x})}{P(y=0|s=1)}.$$

Letting $\mathcal{L}^*(\boldsymbol{\beta}) = \prod_{i=1}^n \pi^*(\boldsymbol{x_i})^{y_i} \cdot [1-\pi^*(\boldsymbol{x_i})]^{1-y_i}$, the case-control likelihood function shown in equation (2.5) becomes

$$\mathcal{L}^*(\boldsymbol{\beta}) \cdot \prod_{i=1}^n \left[\frac{P(\boldsymbol{x_i})}{P(y_i|s_i=1)}\right],$$

where the probability $P(y_i|s_i = 1)$ in the denominator is the study sampling fraction for either cases ($y_i = 1$) or controls ($y_i = 0$), which is fixed by the case-control design, and $\mathcal{L}^*(\boldsymbol{\beta})$ is the likelihood obtained when we pretend that the case-control data were collected in a prospective cohort study. Therefore, under the assumption that the marginal distribution of covariates $\boldsymbol{x}$, $P(\boldsymbol{x})$, does not contain information about the parameters in the logistic regression model, maximization of the full likelihood with respect to the parameters in $\pi^*(\boldsymbol{x})$ only needs to consider the portion of the likelihood which looks like a cohort study, implying that the MLE of the odds-ratio estimators from case-control data can be obtained with logistic regression in the same way as prospective data. Prentice and Pyke (1979) derived the large-sample distribution of this MLE and showed that its variance is also the same.

2.2 Firth Logistic Regression

Maximum likelihood estimates of $\boldsymbol{\theta}$ are biased away from 0, as the expectation of the estimate is always larger in absolute value than the true parameter (Nemes et al., 2009). The bias of the ML estimates of $\boldsymbol{\theta}$ can be expanded asymptotically as

$$Bias(\boldsymbol{\theta}) = \mathrm{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta} = \frac{B_1(\boldsymbol{\theta})}{n} + \frac{B_2(\boldsymbol{\theta})}{n^2} + \cdots .$$

Most bias-corrective methods remove the first asymptotic order bias from $\hat{\boldsymbol{\theta}}$ by using $\hat{\boldsymbol{\theta}}_{BC} = \hat{\boldsymbol{\theta}} - B_1(\hat{\boldsymbol{\theta}})/n$. This kind of method relies on obtaining the MLE and then correcting it by subtracting the first-order bias $B_1(\boldsymbol{\theta})/n$. As it requires the existence of the MLE for the sample, it is not feasible for situations in which there is complete or quasi-complete separation and the MLEs do not exist. To address this problem, Firth (1993) derived a bias-preventive approach in that the parameter is not corrected after it is estimated, but a systematic corrective procedure is applied to the score function from which the parameter estimate is calculated. Firth's method guarantees consistent estimates of logistic regression parameters in the presence of separation (Heinze and Schemper, 2002).

2.2.1 The Approach

The idea behind Firth's approach is to implement a small bias in the score function, which counteracts the first order bias $O(n^{-1})$ of the maximum likelihood estimator. This is usually referred to as a bias-preventive method, and a suitable modification to $U(\boldsymbol{\theta})$ is given by

$$U^*(\boldsymbol{\theta}) = U(\boldsymbol{\theta}) - \mathbf{I}(\boldsymbol{\theta})B_1(\boldsymbol{\theta})/n,$$

where $\mathbf{I}(\boldsymbol{\theta})$ denotes Fisher's information of the sample, defined as the negative expected value of the first derivative of $U(\boldsymbol{\theta})$. The modified score function $U^*(\boldsymbol{\theta})$ originates from the simple triangle geometry shown in Figure 2.1 adapted from Firth (1993). If the MLE $\hat{\boldsymbol{\theta}}$ has a positive first-order bias of $B_1(\boldsymbol{\theta})/n$, it can be removed by shifting the score function downward by $\mathbf{I}(\boldsymbol{\theta})B_1(\boldsymbol{\theta})/n$, where the gradient of $U(\boldsymbol{\theta})$ is given by $\partial U(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = -\mathbf{I}(\theta)$. The corresponding estimate $\boldsymbol{\theta}^*$ can then be calculated by setting the modified score function to 0, i.e. $U^*(\boldsymbol{\theta}) = 0$.

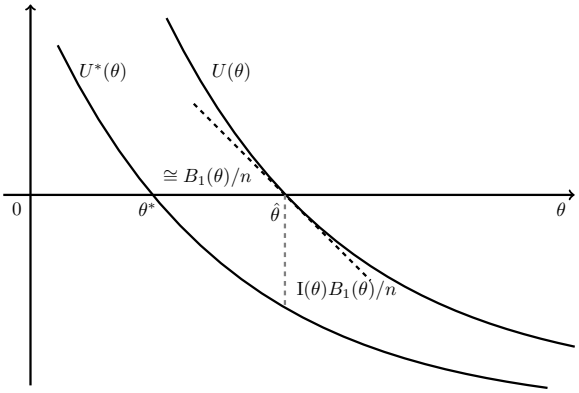


Figure 2.1: Modification of the score function

Firth's approach can also be described as a penalized likelihood method. The usual likelihood function $\mathcal{L}(\boldsymbol{\theta})$ is penalized by a factor equal to the square root of the determinant of the information matrix $|\mathbf{I}(\boldsymbol{\theta})|^{\frac{1}{2}}$. Firth (1993) also showed that if the target parameter is the canonical parameter of an exponential family, his correction scheme is equivalent to penalizing the likelihood by the Jeffery's invariant prior, which is essentially the square root of the log determinant of the Fisher information matrix of the parameters. The penalized likelihood function for Firth's model is thus

$$\mathcal{L}^*(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) \cdot |\mathbf{I}(\boldsymbol{\theta})|^{\frac{1}{2}}. \tag{2.9}$$

Taking the natural logarithm of equation (2.9) yields the corresponding penalized log likelihood function:

$$\ell^*(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \frac{1}{2}\log|\mathbf{I}(\boldsymbol{\theta})|.$$

If Firth's method is applied to the binary logistic regression model defined in equation (2.1), where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})^T$, it is generally known as Firth logistic regression. The resulting penalized log likelihood function is

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n}\{y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)\} + \frac{1}{2}\log|\mathbf{I}(\boldsymbol{\theta})|, \tag{2.10}$$

in which the information matrix is $\mathbf{I}(\boldsymbol{\theta}) = \mathbf{X^T W X}$, with $\mathbf{W} = \text{diag}[\pi_i(1-\pi_i)]$ and $\pi_i = P(y=1|\boldsymbol{x_i}, \boldsymbol{\theta})$. The second term on the right-hand side of equation (2.10) is maximized at $\pi_i = 0.5$ for $i = 1, \ldots, n$, which occurs when $\boldsymbol{\theta} = 0$. Therefore, the parameters are shrunken towards zero; the penalized-likelihood estimates will typically be smaller in absolute value than standard MLEs.

Heinze and Schemper (2002) applied Firth logistic regression to data sets with separation. Their simulation results showed that Firth's penalized likelihood estimator is an ideal solution to the separation problem in logistic regression, as the resultant estimator does not depend on the existence of the classical maximum likelihood estimator. They have extensively compared the estimators from Firth's method to the ordinary MLE in small samples, and have found the Firth method to be consistently superior: point estimates are more precise (i.e. have lower variability), and confidence intervals are more reliable in terms of coverage probabilities. Separately, Bull et al. (2002) extended Firth's method to multinomial logistic regression and found the extension to be superior to other methods in simulation studies involving small samples. They confirmed that Firth's estimator becomes equivalent to the maximum likelihood estimator as sample size increases.

2.2.2 Point and Interval Estimates

Taking the derivative of equation 2.10 with respect to $\boldsymbol{\theta}$, the modified score function has the following form:

$$\frac{\partial \ell^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = U^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[y_i - \pi_i + h_i(\frac{1}{2} - \pi_i) \right] \boldsymbol{x}_i,$$

where h_i is the ith diagonal element of the penalized version of hat matrix:

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}(\mathbf{X^T W X})^{-1}\mathbf{X^T W}^{\frac{1}{2}}.$$

Penalized maximum likelihood estimates can be obtained through application of the standard numerical routine described in equation (2.3) with the $U(\boldsymbol{\theta}^{(r)})$ term replaced by $U^*(\boldsymbol{\theta}^{(r)})$. By imposing the penalty term at each step in the iteration process, this modified score function prevents the estimates from going off to infinity and failing to converge and ensures finite ML estimates when there is separation in the data. Similarly, the standard error can be estimated based on the roots of the diagonal elements of $\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})$, the standard information matrix from the unpenalized log likelihood evaluated at $\hat{\boldsymbol{\theta}}$.

The $(1-\alpha)\times 100\%$ Wald confidence interval for $\boldsymbol{\theta}$ is calculated as $(\hat{\boldsymbol{\theta}} - z_{1-\alpha/2}\sqrt{\mathbf{I}(\hat{\boldsymbol{\theta}})};\ \hat{\boldsymbol{\theta}} + z_{1-\alpha/2}\sqrt{\mathbf{I}(\hat{\boldsymbol{\theta}})})$, which assumes the normal sample distribution of parameter estimates. How-

ever, this assumption is often violated when Firth's approach is used to fit logistic models in datasets where separation exists, as the resulting estimates are typically approaching a boundary of the parameter space and the likelihood profile is asymmetric.

Unlike Wald's method, the profile likelihood based method of constructing confidence intervals allows for asymmetric distributions. Let β_k denote the kth element of $\boldsymbol{\theta}$. The end points, $\beta_{k,\text{lower}}$ and $\beta_{k,\text{upper}}$, for a two-sided $(1-\alpha)\times 100\%$ profile likelihood based interval for β_k, are given by the solution to:

$$2[\ell_p^*(\hat{\beta}_k) - \ell_p^*(\beta_{k,\text{upper}})] = 2[\ell_p^*(\hat{\beta}_k) - \ell_p^*(\beta_{k,\text{lower}})] \sim \chi^2_{1,1-\alpha} \tag{2.11}$$

where $\ell_p^*(\beta_k)$ is the penalized profile log likelihood obtained by fixing β_k and maximizing the penalized likelihood in equation (2.10) over all parameters in $\boldsymbol{\theta}$ other than β_k. A value of the profile log-likelihood is computed by first specifying a value for the coefficient of interest, β_k, and then finding the value of the other coefficients that maximizes the log-likelihood. This process is repeated over a grid of values of the specified coefficients, until the solutions to equation (2.11) are found. A $(1-\alpha)\times 100\%$ profile likelihood confidence interval for β_k is the continuous set of values β_k for which twice the difference of the maximized log likelihood and the profile likelihood at β_k does not exceed the $(1-\alpha)\times 100$ percentile of the χ_1^2 -distribution.

It is widely known that, in small samples, likelihood-based confidence intervals tend to provide more accurate coverage than Wald confidence intervals. Heinze and Schemper (2002) have shown that the Firth penalized profile likelihoods for the coefficients are often asymmetrical and thus the inference based on Wald statistics can be misleading. They have shown that the profile-likelihood based confidence interval has better coverage properties than the symmetric Wald confidence interval, and have recommended the former interval estimator for Firth logistic regression.

2.3 Zhang Logistic Regression

Based on Firth's approach, Zhang (2006) proposed a Firth-like preventive approach to reduce the first order bias of ML estimates under the logistic model based on case-control data. In analogy to Firth's method, where a small bias term is introduced to modify the standard score function, Zhang's method modifies the score function arising from a profile likelihood that is derived under a two-sample semiparametric model for case-control data.

2.3.1 A Two-sample Semiparametric Model

As we have discussed in section 2.1, the standard logistic regression model is typically used in prospective studies; i.e., we observe the covariate value of each individual first and then measure the binary response. Case-control data arises from retrospective sampling, in which we have the binary response first and then observe their covariates. In other words, we have one sample from the control population, and another sample from the case population.

Let $X_1, \ldots, X_{n_0}$ be an independent and identically distributed (i.i.d.) sample from the control population ($y = 0$), and $X_{n_0+1}, \ldots, X_n$ be another i.i.d sample from the case population ($y = 1$). Let $n = n_0 + n_1$ be the combined sample size, $\phi = P(y = 1) = 1 - P(y = 0)$, and $P(\boldsymbol{x})$ be the marginal distribution of $\boldsymbol{x}$. Then, applying conditional probability calculations, we can write

$$\begin{aligned} P(\boldsymbol{x}|y = 1) &= \frac{P(y = 1|\boldsymbol{x}) \cdot P(\boldsymbol{x})}{P(y = 1)} = \frac{\pi_i}{\phi} \cdot P(\boldsymbol{x}), \text{ and similarly,} \\ P(\boldsymbol{x}|y = 0) &= \frac{P(y = 0|\boldsymbol{x}) \cdot P(\boldsymbol{x})}{P(y = 0)} = \frac{1 - \pi_i}{1 - \phi} \cdot P(\boldsymbol{x}), \end{aligned}$$

where π_i stands for the standard logistic regression model defined in equation (2.1). Taking the ratio of the two equations, we have

$$\frac{P(\boldsymbol{x}|y = 1)}{P(\boldsymbol{x}|y = 0)} = \frac{\pi_i}{1 - \pi_i} \cdot \frac{1 - \phi}{\phi}.$$

Letting $g(\boldsymbol{x})$ denote the conditional density function $P(\boldsymbol{x}|y = 0)$, and $h(\boldsymbol{x})$ be the conditional density function $P(\boldsymbol{x}|y = 1)$, we can rewrite the previous formula as

$$\begin{aligned} h(\boldsymbol{x}) = P(\boldsymbol{x}|y = 1) &= \frac{\pi_i}{1 - \pi_i} \cdot \frac{1 - \phi}{\phi} \cdot g(\boldsymbol{x}) \\ &= \exp(\alpha + \boldsymbol{x}^T\boldsymbol{\beta}) \cdot \frac{1 - \phi}{\phi} \cdot g(\boldsymbol{x}) \\ &= \exp\left[\alpha + \log\left(\frac{1 - \phi}{\phi}\right) + \boldsymbol{x}^T\boldsymbol{\beta}\right] \cdot g(\boldsymbol{x}) \\ &= c(\boldsymbol{\beta}, g) \exp(\boldsymbol{x}^T\boldsymbol{\beta}) \cdot g(\boldsymbol{x}), \end{aligned} \tag{2.12}$$

where $c(\boldsymbol{\beta}, g) = \exp\left[\alpha + \log\left(\frac{1-\phi}{\phi}\right)\right]$. Note that $h(\boldsymbol{x})$ is a density function and so it has to integrate to 1 over the values of $\boldsymbol{x}$. Thus, $c(\boldsymbol{\beta}, g) = \exp\left[\alpha + \log\left(\frac{1-\phi}{\phi}\right)\right]$ is equal to the inverse of the integral of $\exp(\boldsymbol{x}^T\boldsymbol{\beta})g(\boldsymbol{x})$ with respect to $\boldsymbol{x}$. Therefore, $c(\boldsymbol{\beta}, g)$ is a function of $(\boldsymbol{\beta}, g)$.

As a result, we have a two-sample semiparametric model in which there are two independent samples from

$$\begin{aligned} &P(\boldsymbol{x}|y=0) = g(\boldsymbol{x}), \quad \text{and} \\ &P(\boldsymbol{x}|y=1) = h(\boldsymbol{x}) = c(\boldsymbol{\beta}, g)\exp(\boldsymbol{x}^T\boldsymbol{\beta}) \cdot g(\boldsymbol{x}). \end{aligned} \tag{2.13}$$

Qin and Zhang (1997) show that the prospective logistic regression model (2.1) and the retrospective two-sample semiparametric model (2.13) are equivalent with parameters related by

$$\log c(\boldsymbol{\beta}, g) = \alpha + \log\left(\frac{1-\phi}{\phi}\right). \tag{2.14}$$

Model (2.13) is a semiparametric model because it has both finite dimensional unknown parameter of interest $\boldsymbol{\beta}$, and an infinite dimensional unknown distribution function g. Both of the parameters $\boldsymbol{\beta}$ and the density g of the covariates in the control group are to be estimated.

2.3.2 Profile Likelihood Function

Based on the model defined in (2.13), the full likelihood function for a sample with n_0 controls and $n_1 = n - n_0$ cases can be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, g) &= \prod_{i=1}^{n_0} P(\boldsymbol{x_i}|y=0) \cdot \prod_{i=n_0+1}^{n} P(\boldsymbol{x_i}|y=1) \\ &= \prod_{i=1}^{n_0} g(\boldsymbol{x_i}) \cdot \prod_{i=n_0+1}^{n} c(\boldsymbol{\beta}, g)\exp(\boldsymbol{x_i}^T\boldsymbol{\beta}) g(\boldsymbol{x_i}) \\ &= \prod_{i=1}^{n} g(\boldsymbol{x_i}) \cdot \prod_{i=n_0+1}^{n} \exp[\log(c(\boldsymbol{\beta}, g)) + \boldsymbol{x_i}^T\boldsymbol{\beta}]. \end{aligned} \tag{2.15}$$

Combining equation (2.8) and (2.14), we can rewrite the expression of $\log c(\boldsymbol{\beta}, g)$ as $\log c(\boldsymbol{\beta}, g) = \alpha^* - \log(n_1/n_0)$. By doing this, we can overparametrize the case-control likelihood (2.15) by including α^* as a parameter. As a result, equation (2.15) becomes

$$\begin{aligned} \mathcal{L}(\alpha^*, \boldsymbol{\beta}, g) &= \prod_{i=1}^{n} g(\boldsymbol{x_i}) \cdot \prod_{i=n_0+1}^{n} \exp[\alpha^* - \log(n_1/n_0) + \boldsymbol{x_i}^T\boldsymbol{\beta}] \\ &= \prod_{i=1}^{n} g(\boldsymbol{x_i}) \cdot \prod_{i=n_0+1}^{n} w(\boldsymbol{x_i}), \end{aligned} \tag{2.16}$$

where $w(\boldsymbol{x_i}) = \exp[\alpha^* - \log(n_1/n_0) + \boldsymbol{x_i}^T\boldsymbol{\beta}]$.

Qin and Zhang (1997) showed that a profile likelihood function can be obtained by maximizing (2.16) over the nonparametric nuisance parameter g for fixed $(\alpha^*, \boldsymbol{\beta})$, subject to

$$g(\boldsymbol{x_i}) \geq 0, \ \sum g(\boldsymbol{x_i}) = 1, \ \sum g(\boldsymbol{x_i})[w(\boldsymbol{x_i}) - 1] = 0.$$

The first and second constraints ensure that the $g(\boldsymbol{x_i})$'s comprise a probability distribution function for the controls, i.e. nonnegative everywhere and sum to one. The third constraint reflects the fact that the set of $h(\boldsymbol{x_i}) = w(\boldsymbol{x_i})g(\boldsymbol{x_i})$ values $(i = 1, \ldots, n)$ comprise a distribution function for the cases.

Using the approach of Qin and Lawless (1994), with fixed $(\alpha^*, \boldsymbol{\beta})$, the maximum value of the two sample case control likelihood function (2.16) is obtained at

$$\widetilde{g}(\boldsymbol{x_i}) = \frac{1}{n_0[1 + \frac{n_1}{n_0} \exp(\alpha^* - \log(n_1/n_0) + x_i\boldsymbol{\beta})]} = \frac{1}{n_0[1 + \exp(\alpha^* + x_i\boldsymbol{\beta})]}.$$

Inserting $\widetilde{g}(\boldsymbol{x_i})$ back into equation (2.16), we obtain the semiparametric profile log likelihood function as a function of $(\alpha^*, \boldsymbol{\beta})$ only, ignoring a constant:

$$l(\alpha^*, \boldsymbol{\beta}) = \sum_{j=n_0+1}^{n} (\alpha^* + x_j\boldsymbol{\beta}) - \sum_{i=1}^{n} \log(1 + \exp(\alpha^* + x_i\boldsymbol{\beta})). \tag{2.17}$$

2.3.3 Zhang Method

Zhang's approach is an extension of Firth's method to case-control data by introducing a small bias term into the score function obtained from the semiparametric profile log likelihood in equation (2.17). Let $\boldsymbol{\theta^*} = (\alpha^*, \boldsymbol{\beta})^T$. Denote $U_p(\boldsymbol{\theta^*})$ as the corresponding semiparametric profile score function, which is obtained by taking the derivative of equation (2.17). Zhang (2006) employs a Firth-like preventive approach to bias reduction by imposing a penalty term on $U_p(\boldsymbol{\theta^*})$, producing a modified semiparametric profile score function $U_p^*(\boldsymbol{\theta^*})$,

$$U_p^*(\boldsymbol{\theta^*}) = U_p(\boldsymbol{\theta^*}) - \widetilde{\mathbf{I}}(\boldsymbol{\theta^*})\tilde{B}_1/n,$$

where $\widetilde{\mathbf{I}}(\boldsymbol{\theta^*})$ and $\tilde{B}_1$ are the empirical information matrix and bias terms based on the profile log likelihood function (2.17), respectively. Zhang (2006) gives the explicit formulae for $\widetilde{\mathbf{I}}(\boldsymbol{\theta^*})$ and $\tilde{B}_1$.

The estimator is the root of $U_p^*(\boldsymbol{\theta^*}) = 0$, and can be obtained using the iterative method described in the previous section. Both the Wald and profile likelihood based confidence intervals can be obtained the same way as described in the section 2.2.2. We refer to this method as Zhang logistic regression.

Chapter 3

Simulation Study

Firth logistic regression has been widely applied to data from case-control studies of rare disease, as it allows for convergence to finite estimates in conditions of separation. Also, the implementation of Firth logistic regression is fairly easy as it is available in many standard statistical software environments including R, SAS, and Stata. However, researchers often neglect the fact that Firth logistic regression has been developed under a prospective sampling design, and its application in retrospective case-control data has not yet been formally justified. A simulation study with large samples (Ma et al., 2013) gave promising results showing that Firth logistic regression analysis controls Type I error for both balanced and unbalanced case-control data.

Unlike Firth's approach which is developed for prospective data, Zhang's method is derived directly from the case-control likelihood, and has been demonstrated to remove the first order bias. Zhang (2006) has shown that, under equal numbers of cases and controls, the point estimators from Firth logistic regression are the same as those from his profile-likelihood approach for case-control data. Therefore, the purpose of the simulation study is two-fold: (1) to confirm that the point estimators from Firth and Zhang logistic regression are the same when the number of cases and controls is balanced in the study design; and (2) to assess the performance of the point and interval estimators from the two approaches when the case-control study design is not balanced.

3.1 Design of Study

The case-control data were simulated from the two-sample semiparametric model (2.13). For simplicity, we incorporated a single continuous covariate x. The main parameter of interest, β, is the log odds ratio associated with x. In the simulation study, we assume

that the covariate distribution in the control group follows a standard normal distribution, $g(x) \sim N(0,1)$. From equation (2.12), the covariate distribution of the case group $h(x)$ is then

$$h(x) \propto \exp(x\beta) \cdot g(x) = \exp(x\beta) \cdot \exp\left(-\frac{x^2}{2}\right) = \exp\left[-\frac{1}{2}(x-\beta)^2\right] \cdot \exp\left(\frac{\beta^2}{2}\right),$$

and so the density function for cases is a $N(\beta, 1)$ distribution.

Heinze and Schemper (2002) showed that the probability of separation depends on the magnitude of the odds ratios, sample size, and the degree of balance between cases and controls. In the simulation studies, we consider different scenarios by varying these three settings. The true β was set to 0, 1, 2, 3, 4, or 5, corresponding to odds ratio of 1, 2.72, 7.39, 20.09, 54.60, and 148.41 respectively. We consider a small sample $n = 50$ and a moderate sample $n = 100$. For each sample size, we consider both the balanced design ($n_0 = n_1$) and the most commonly used unbalanced study ($n_0/n_1 = 4:1$). Therefore, for each pair of (n_0, n_1, β), we generated 1000 independent data sets of combined random samples from the $N(0,1)$ and $N(\beta,1)$ populations with

$$x_1, \ldots, x_{n_0} \sim N(0,1); \quad x_{n_0+1}, \ldots, x_{n_0+n_1} \sim N(\beta, 1).$$

Separation in these generated data sets was determined by fitting a standard binary logistic regression model with `glm()` in `R` and checking whether the maximum likelihood estimation of parameters converged. The number of data sets with separation (i.e., when maximum likelihood estimation with `glm()` gave the warning message "`fitted probabilities numerically 0 or 1 occurred.`") was recorded.

For data sets without separation, we obtained the uncorrected MLE of the log odds ratio, $\hat{\beta}_{\text{MLE}}$, and the penalized MLEs, $\hat{\beta}_{\text{Firth}}$ and $\hat{\beta}_{\text{Zhang}}$, from Firth logistic regression and Zhang logistic regression, respectively. For data sets with separation, only the Firth estimates and the Zhang estimates were obtained, as MLEs do not exist. MLEs were obtained from applying the `glm()` function in `R` with binomial link. Firth logistic regression estimates were obtained from applying the `logistf()` function in the `logistf` package (Heinze et al., 2013), whereas Zhang logistic regression estimates were obtained from `R` functions we implemented ourselves (see Appendix for code).

Based on the $B = 1000$ estimates from Monte Carlo simulation, we evaluated the biases, variances, and mean square error (MSE) of the point estimators. Let β be the true value of a regression parameter. Let $\hat{\beta}$ be the estimator of β and $\hat{\beta}^{(r)}$ be its realization in the rth

simulation replicate. The estimated bias of $\hat{\beta}$ is:

$$\widehat{bias(\hat{\beta})} = \bar{\hat{\beta}} - \beta = \frac{1}{B}\sum_{r=1}^{B}(\hat{\beta}^{(r)} - \beta).$$

A positive bias suggests overestimation of the covariate effect, and a negative bias indicates underestimation. The empirical variance of $\hat{\beta}$ is:

$$\widehat{\mathrm{Var}(\hat{\beta})} = \frac{1}{B}\sum_{r=1}^{B}(\hat{\beta}^{(r)} - \bar{\hat{\beta}})^2.$$

The MSE can then be expressed as $MSE = bias(\hat{\beta})^2 + \mathrm{Var}(\hat{\beta})$, which provides an overall measurement of precision and accuracy. The calculation of bias and variance of the MLE were only based on unseparated data sets.

We also considered two sets of interval estimators, $(1-\alpha)\times 100\%$ Wald confidence intervals and profile likelihood based confidence intervals, and compared their coverage probability and length. The coverage probability (CP) measures the percentage of times the confidence interval catches the true point estimate in the 1000 simulated data sets:

$$\widehat{\mathrm{CP}} = \frac{1}{B}\sum_{r=1}^{B} I\{\hat{\beta}^{(r)} \in [\mathrm{CI}_{lower}, \mathrm{CI}_{upper}]\},$$

where I is an indicator function, and CI_{lower} and CI_{upper} denote the lower and upper bounds of the corresponding confidence interval, respectively.

3.2 Results

Table 3.1: The frequency of separation in the 1000 simulated data sets.

n	n_0	n_1	$\beta=0$	$\beta=1$	$\beta=2$	$\beta=3$	$\beta=4$	$\beta=5$
50	25	25	0	0	4	139	654	942
	40	10	0	0	15	247	729	960
100	50	50	0	0	0	21	361	826
	80	20	0	0	1	49	470	878

Table 3.1 presents the frequency of separation in the simulated data sets. Each entry is based on 1000 samples. In agreement with Heinze and Schemper (2002), separation is more likely in the small, unbalanced samples, with extreme log odds ratios. As shown in Table 3.1, when $\beta = 0$ or $\beta = 1$, separation occurs in none of the generated data sets. When the log odds ratio increases to $\beta = 5$, even with a moderate sample size ($n = 100$) and balanced

design ($n_0 = n_1$), 82.6% of the data sets are generated with separation and so the MLE does not exist.

The Firth and Zhang estimates of the log-odds ratio are obtained for all the data sets, but MLEs are obtained only for the data sets without separation. Generally, the biases of the point estimators vary depending on the values of ($n_0,\ n_1,\ \beta$). For the settings where the log odds ratio $\beta = 0,\ 1,\ 2,\ 3$ and 4, we present full results on bias, variance, and MSE of the point estimators in Table 3.2, and detailed coverage comparison of the corresponding interval estimators in Table 3.3. As when $\beta = 5$, separation exists in more than 80% of the generated data sets, we only evaluate penalized estimators, and the results are summarized in Table 3.4.

3.2.1 Evaluation of Point Estimators

Figures 3.1 and 3.2 present the distribution of the Firth estimates and the corresponding bias for the simulation scenarios with 10 cases and 40 controls, and 50 cases and 50 controls, respectively. By inspection of the figures and Table 3.2, we observe that:

(1) Overall, the MLEs have finite sample bias that is proportional to the true parameter value and increases in magnitude as the sample size decreases. In most settings ($\beta \leq 3$), both Firth estimates and Zhang estimates are closer to the true parameter than the MLEs, and could be obtained in all samples, even when there are infinite MLEs. Also, both Zhang and Firth logistic regression give very similar point estimates and variance across replications for all the simulation scenarios, even for unbalanced case-control data sets with 4:1 control to case ratio, as the two sets of density distributions in Figure 3.3 are indistinguishable.

(2) As expected, the point estimators are all unbiased when the true β is zero. As β increases, the Firth and Zhang estimators tend to be smaller than the MLEs, as expected from the bias reducing property of these penalized likelihood methods, but are comparable to each other. For $\beta = 1,\ 2,\ 3$, the bias of the Firth and Zhang estimators is close to zero, and finite sample bias is effectively eliminated. When β increases to 4 and 5, both the Firth and Zhang estimators overcorrect, as the bias of these estimators is shifted downwards, away from 0. As the bias increases, the variance of the Firth and Zhang estimators decreases. From Figure 3.3, the distribution of the estimator becomes increasingly skewed to the right as β increases.

(3) It is also worth noting that, when β gets large ($\beta = 4$ or 5), there is a substantial proportion of the data sets with separation; therefore the calculation of bias, variance and MSE for MLEs is only based on a small fraction of the 1000 generated data sets. In

all the simulation runs, the variance of the Firth and Zhang estimators is consistently less than that of MLEs; this suggests that MLEs are not stable when β is large.

(4) As the sample size increases from small to moderate, or the design changes from unbalanced to balanced, the magnitude of the bias of the Firth and Zhang estimators is smaller and the variability in the estimators is less.

Since MLEs are not available in data sets with separation, we are also curious about whether the Firth estimator behaves differently in data sets with or without separation. Figure 3.4 shows the distribution of the Firth estimators in the scenario where $n_0 = 25,\ n_1 = 25,\ \beta = 4$. In this scenario, separation occurs in 64.5% of the data sets. Comparison of the distributions of the Firth estimator in data sets with or without separation indicates that data sets with separation tend to yield estimates with larger values. Separation is common with large values of beta and so datasets without separation tend to be more consistent with values of beta that are smaller. The fact that the datasets without separation tend to be more compatible with smaller values of beta than the values under which they were generated implies that the Firth estimators tends to overcorrect the bias for these data sets.

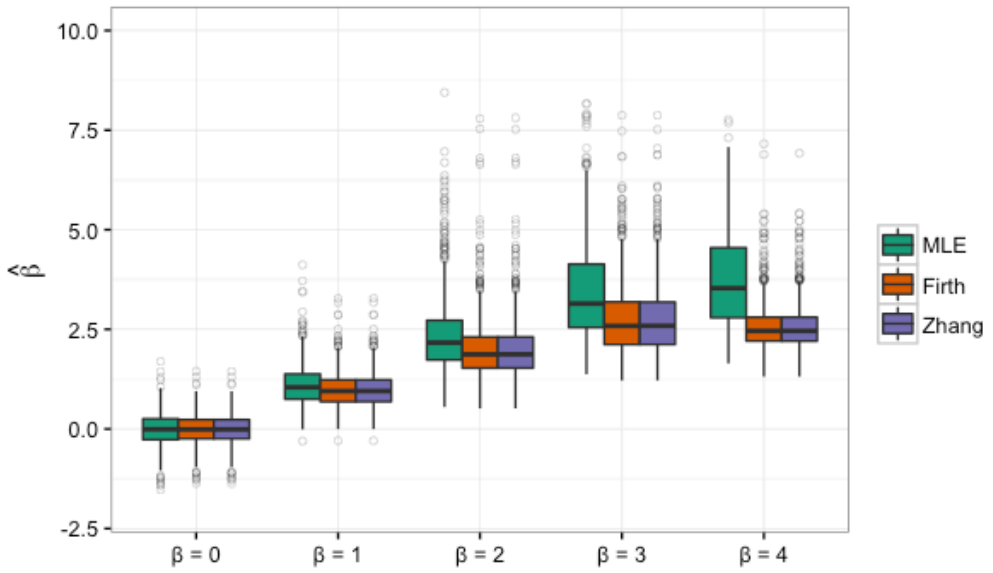
3.2.2 Evaluation of Interval Estimators

As shown in Table 3.3, when $\beta = 0$ or 1, the median lengths of the Firth and Zhang interval estimators are comparable to those of the likelihood-based interval estimators. When $\beta \geq 2$, the median lengths of both penalized interval estimators are shorter than those of the MLEs, for both Wald and profile confidence intervals, although the median interval lengths for all methods increase when there are data sets with separation. It is also worth noting that, when β gets large ($\beta = 4$), there is a substantial proportion of the data sets although not meeting the criteria for separation, have infinity upper limit for profile based confidence intervals for MLEs, resulting in infinity interval length and higher than nominal coverage probability.

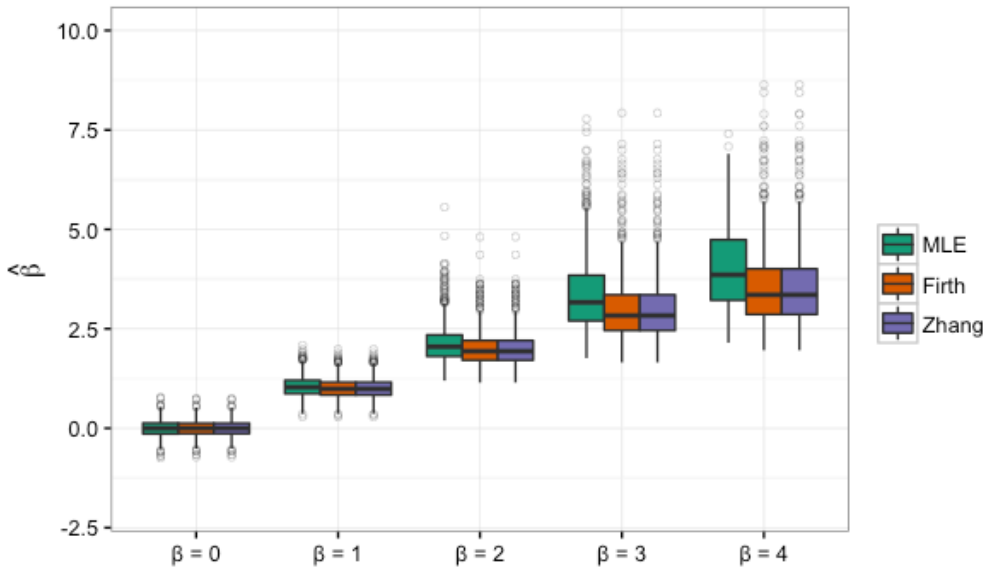
The coverage probabilities of the Wald and profile likelihood based confidence intervals for the Firth and Zhang estimators are virtually identical. For low odds ratios ($\beta = 0,\ 1,\ 2$) where separation is not likely to occur, the empirical coverage of both the profile-penalized-likelihood- and Wald-based confidence intervals is equally satisfactory. The performance of Wald confidence intervals is particularly sensitive to the phenomenon of separation, whereas the profile intervals are somewhat less so. This can be explained by the symmetric construction of Wald confidence intervals. For data sets with separation, the profile of the penalized likelihood function is highly asymmetric. The inappropriate symmetry of the Wald confidence intervals reflected in the coverage probabilities substantially departing from 95% in situations with a high probability of separation (high odds ratio, small sample, unbalanced

design). In these cases coverage by penalized profile likelihood confidence intervals is much more satisfactory, achieving close to nominal coverage for log odds ratios as large as 4, even in small, unbalanced data sets (10 cases, 40 controls). However, for extreme log odds ratios such as $\beta = 5$, the coverage probability of profile likelihood based confidence intervals is below nominal as well (see Table 3.4) .

In most cases, the length of the confidence intervals increases as the true value of β increases. When $\beta \neq 0$, penalized profile likelihood confidence intervals have greater length than the Wald confidence intervals. Also, it is interesting to note that penalized profile likelihood confidence intervals for both Firth and Zhang estimators have comparable lengths when $\beta \leq 3$ but, when $\beta \geq 4$, the median length of the Zhang interval estimator is longer than the median length of the Firth estimator. As shown in Table 3.4, when $\beta = 5$, both the Firth and Zhang penalized profile likelihood interval estimators have below-nominal coverage. However, the coverage of Zhang's interval is closer to nominal than Firth's, perhaps due to its longer interval length. From Figure 3.5, in data sets without separation, the distributions of the length of the Firth and Zhang penalized profile likelihood intervals are virtually identical, but, in data sets with separation, the Firth intervals tend to be shorter. Finally, profile likelihood based confidence intervals for both Firth and Zhang estimators tend to be longer in data sets with separation than in data sets without separation.

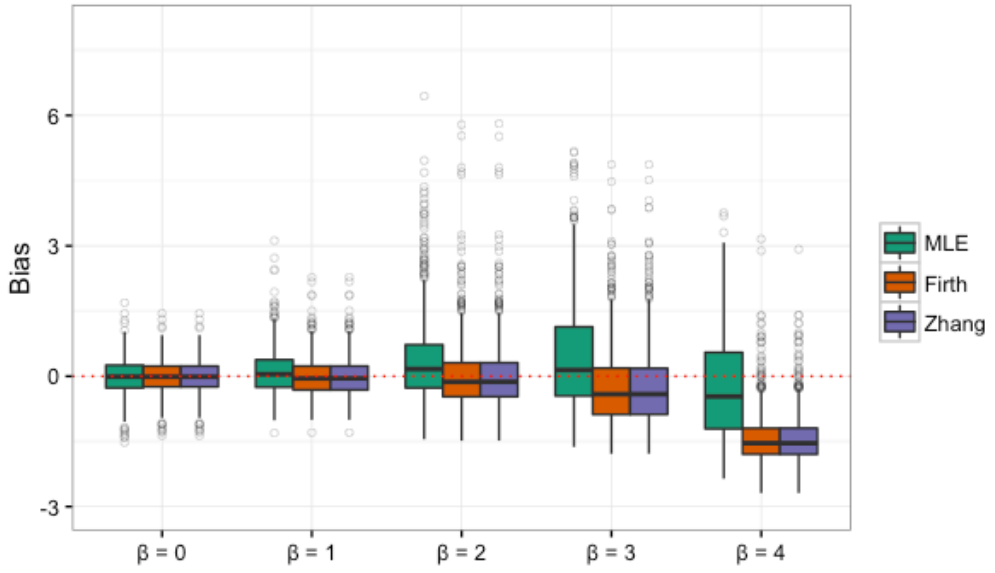


(a)

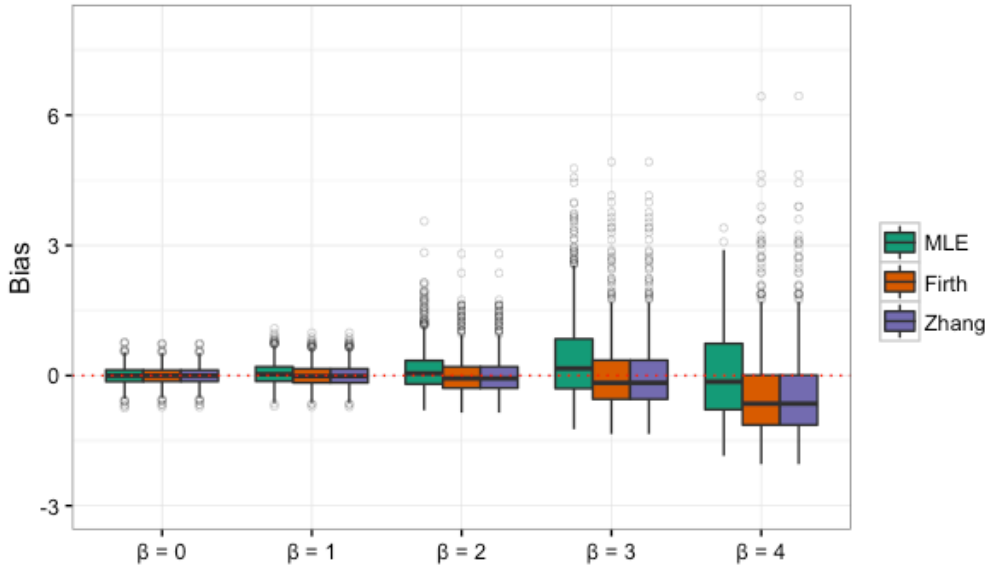


(b)

Figure 3.1: Boxplots of the distribution of MLEs, Firth and Zhang estimators under scenarios with (a) 10 cases and 40 controls , and (b) 50 cases and 50 controls. Boxplots are obtained from 1000 simulated data sets as described in the text. For each β value, the estimates distribution of MLEs, Firth, and Zhang estimators are ordered from left to right.

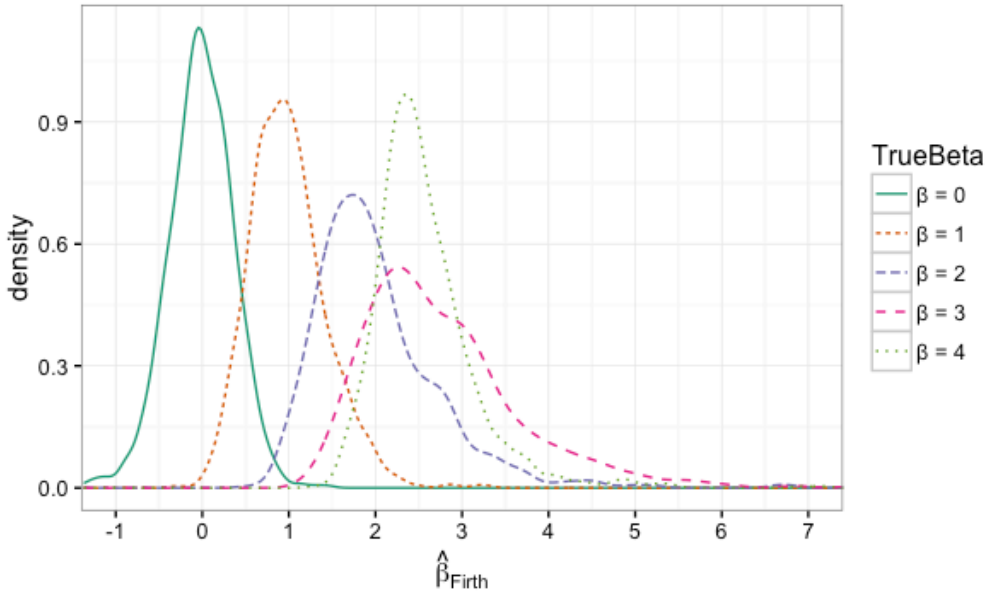


(a)

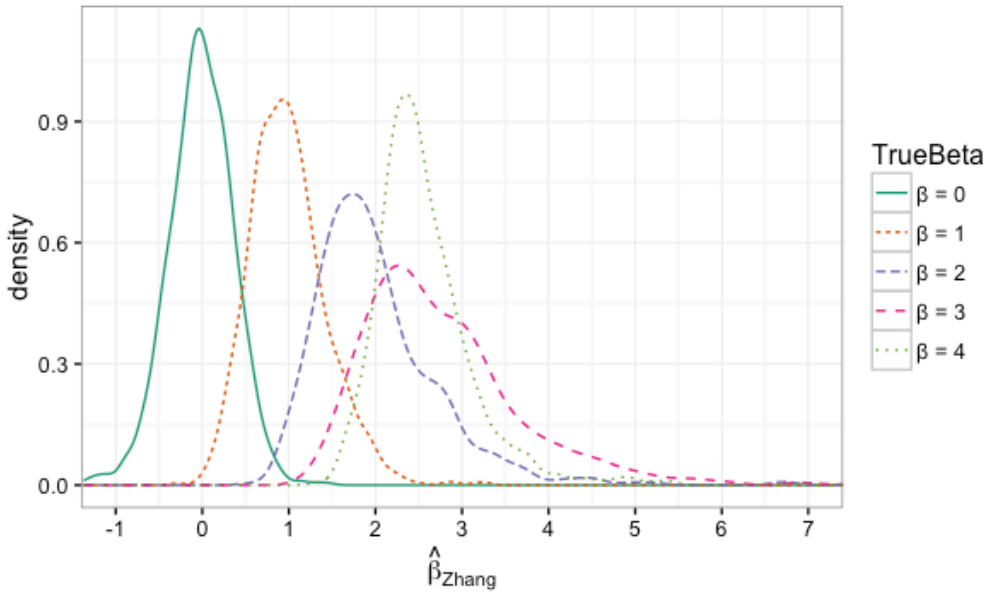


(b)

Figure 3.2: Boxplots of the bias distribution of MLEs, Firth and Zhang estimators under scenarios with (a) 10 cases and 40 controls , and (b) 50 cases and 50 controls. Boxplots are obtained from 1000 simulated datasets as explained in the text. For each β value, the bias distribution of MLEs, Firth, and Zhang estimators are ordered from left to right.



(a)



(b)

Figure 3.3: Density estimates of the distribution of (a) Firth logistic regression estimator $\hat{\beta}_{\text{Firth}}$, and (b) Zhang logistic regression estimator $\hat{\beta}_{\text{Zhang}}$, from simulations of 1000 data sets of 10 cases and 40 controls.

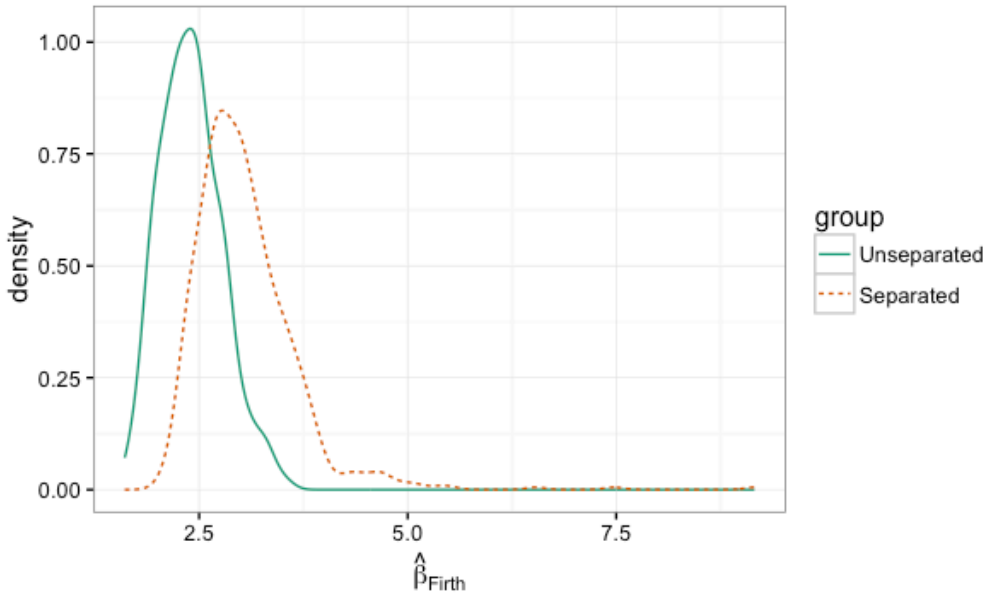


Figure 3.4: Distribution of Firth estimator for data sets with and without separation. The histogram is obtained from 1000 simulated data sets of $n_0 = n_1 = 25$ cases and controls, with log odds ratio $\beta = 4$.

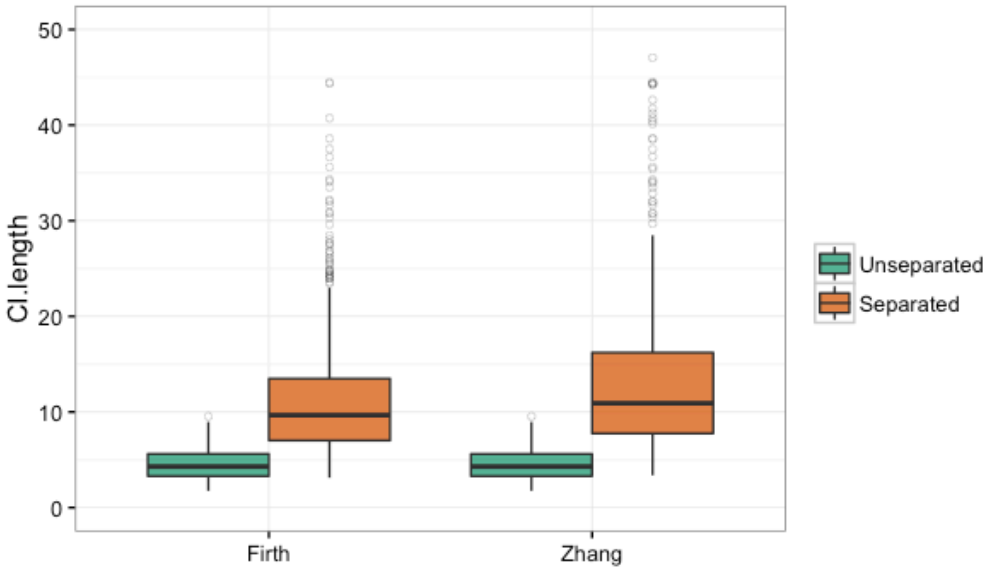


Figure 3.5: Distribution of the length of confidence intervals for data sets with and without separation. The histogram is obtained from 1000 simulated data sets of $n_0 = n_1 = 25$ cases and controls, with log odds ratio $\beta = 4$. The distribution for data sets with separation is on the right, for both the Firth and Zhang estimators

Table 3.2: Operating characteristics of the point estimators based on 1000 simulated data sets at each setting

n	n_0	n_1			$\beta = 0$	$\beta = 1$	$\beta = 2$	$\beta = 3$	$\beta = 4$
50	25	25	Bias	MLE [a]	-0.0187	0.0814	0.2839	0.4533	-0.1578
				Firth	-0.0174	-0.0053	0.0095	-0.1098	-1.1736
				Zhang	-0.0174	-0.0053	0.0095	-0.1097	-1.1730
			Variance	MLE	0.0914	0.1436	0.6812	1.4652	1.2306
				Firth	0.0786	0.1163	0.4704	1.0599	0.3920
				Zhang	0.0786	0.1163	0.4704	1.0611	0.3973
			MSE	MLE	0.0918	0.1502	0.7618	1.6708	1.2555
				Firth	0.0790	0.1163	0.4705	1.0720	1.7694
				Zhang	0.0790	0.1163	0.4705	1.0731	1.7732
	40	10	Bias	MLE	-0.0143	0.1032	0.3768	0.4388	-0.2036
				Firth	-0.0131	-0.0101	0.0268	-0.2385	-1.4334
				Zhang	-0.0131	-0.0092	0.0276	-0.2358	-1.4358
			Variance	MLE	0.1652	0.2619	0.9144	1.5685	1.6546
				Firth	0.1372	0.1938	0.7060	0.8194	0.3436
				Zhang	0.1376	0.1939	0.7094	0.8338	0.3274
			MSE	MLE	0.1654	0.2726	1.0563	1.7610	1.6960
				Firth	0.1373	0.1939	0.7067	0.8763	2.3982
				Zhang	0.1378	0.1940	0.7102	0.8894	2.3889
100	50	50	Bias	MLE	-0.0030	0.0482	0.1339	0.3758	0.0690
				Firth	-0.0029	0.0060	0.0058	0.0116	-0.4506
				Zhang	-0.0029	0.0060	0.0058	0.0116	-0.4503
			Variance	MLE	0.0448	0.0665	0.2472	0.9115	1.1590
				Firth	0.0415	0.0603	0.1932	0.7894	0.9636
				Zhang	0.0415	0.0603	0.1932	0.7898	0.9664
			MSE	MLE	0.0448	0.0688	0.2652	1.0527	1.1638
				Firth	0.0414	0.0604	0.1933	0.7895	1.1667
				Zhang	0.0414	0.0604	0.1933	0.7899	1.1692
	80	20	Bias	MLE	-0.0098	-0.0597	0.1642	0.4082	-0.1142
				Firth	-0.0092	0.0070	0.0030	-0.0278	-0.6842
				Zhang	-0.0092	0.0072	0.0031	0.0315	-0.6347
			Variance	MLE	0.0711	0.0933	0.3068	1.1011	0.9781
				Firth	0.0652	0.0815	0.3084	0.8206	0.7410
				Zhang	0.0653	0.0815	0.3079	0.8209	0.7482
			MSE	MLE	0.0712	0.0969	0.3338	1.2677	0.9911
				Firth	0.0653	0.0815	0.3085	0.8214	1.2091
				Zhang	0.0653	0.0815	0.3079	0.8216	1.2142

[a] The bias, variance and MSE calculations for the MLE are based on data sets without separation only. See Table 3.1 for the numbers of data sets with separation at each setting of the simulation study.

Table 3.3: Coverage probabilities and median lengths (in parentheses) for 95% confidence intervals based on 1000 simulated data sets at each setting

n	n_0	n_1			$\beta = 0$	$\beta = 1$	$\beta = 2$	$\beta = 3$	$\beta = 4$
50	25	25		MLE [a]	0.966 (1.152)	0.966 (1.404)	0.962 (2.312)	0.952 (4.196)	0.913 (6.688)
			Wald	Firth	0.974 (1.160)	0.966 (1.404)	0.924 (2.035)	0.899 (3.232)	0.761 (3.915)
				Zhang	0.974 (1.160)	0.966 (1.355)	0.924 (2.035)	0.899 (3.232)	0.761 (3.915)
				MLE	0.947 (1.168)	0.957 (1.412)	0.935 (2.280)	0.974 (4.869)	0.982 (∞)
			PL [b]	Firth	0.956 (1.142)	0.961 (1.367)	0.938 (2.145)	0.946 (3.840)	0.947 (7.245)
				Zhang	0.956 (1.142)	0.961 (1.367)	0.938 (2.146)	0.946 (3.841)	0.947 (7.769)
	40	10		MLE	0.956 (1.448)	0.961 (1.684)	0.956 (2.697)	0.956 (4.735)	0.904 (6.881)
			Wald	Firth	0.967 (1.428)	0.954 (1.604)	0.936 (2.309)	0.875 (3.546)	0.630 (3.627)
				Zhang	0.967 (1.443)	0.958 (1.611)	0.937 (2.315)	0.875 (3.548)	0.633 (3.630)
				MLE	0.944 (1.474)	0.947 (1.705)	0.932 (2.709)	0.981 (∞ [c])	0.978 (∞)
			PL	Firth	0.951 (1.428)	0.951 (1.625)	0.946 (2.476)	0.951 (4.625)	0.945 (6.800)
				Zhang	0.950 (1.429)	0.951 (1.625)	0.946 (2.476)	0.951 (4.644)	0.945 (7.910)
100	50	50		MLE	0.956 (0.797)	0.955 (0.984)	0.955 (1.553)	0.973 (2.940)	0.951 (5.017)
			Wald	Firth	0.968 (0.803)	0.963 (0.969)	0.936 (1.478)	0.921 (2.485)	0.830 (3.961)
				Zhang	0.968 (0.803)	0.963 (0.969)	0.936 (1.478)	0.921 (2.486)	0.830 (3.961)
				MLE	0.950 (0.804)	0.947 (0.987)	0.933 (1.554)	0.957 (2.958)	0.976 (∞)
			PL	Firth	0.955 (0.796)	0.954 (0.972)	0.942 (1.514)	0.952 (2.698)	0.943 (5.511)
				Zhang	0.955 (0.796)	0.954 (0.972)	0.942 (1.514)	0.952 (2.698)	0.943 (5.522)
	80	20		MLE	0.956(0.998)	0.970 (1.181)	0.961 (1.843)	0.973 (3.338)	0.893 (5.275)
			Wald	Firth	0.965 (0.995)	0.973 (1.154)	0.944 (1.699)	0.921 (2.714)	0.807 (4.043)
				Zhang	0.965 (1.000)	0.973 (1.158)	0.944 (1.703)	0.921 (2.716)	0.807 (4.042)
				MLE	0.949 (1.007)	0.960 (1.188)	0.940 (1.830)	0.970 (3.472)	0.989 (∞)
			PL	Firth	0.953 (0.995)	0.965 (1.164)	0.953 (1.764)	0.960 (3.027)	0.947 (6.307)
				Zhang	0.953 (0.996)	0.965 (1.164)	0.953 (1.764)	0.960 (3.027)	0.947 (6.377)

[a] Results for the MLE are for data sets without separation only. See Table 3.1 for the numbers of data sets with separation at each setting of the simulation study.

[b] Profile-likelihood-based interval estimators

[c] More than half of the confidence interval lengths have infinity upper bounds

Table 3.4: Operating characteristics of point and 95% interval estimators when $\beta = 5$

(n_0, n_1)	Method	Bias	Variance	MSE	Wald		PPL[a]	
					Coverage	Length	Coverage	Length
(25, 25)	Firth	-2.871	0.077	8.321	0.022	2.760	0.640	4.459
	Zhang	-2.871	0.077	8.321	0.022	2.760	0.760	5.668
(40, 10)	Firth	-3.075	0.063	9.519	0.003	2.592	0.496	4.027
	Zhang	-3.075	0.063	9.519	0.005	2.596	0.669	5.004
(50, 50)	Firth	-2.140	0.176	4.754	0.287	3.487	0.911	6.392
	Zhang	-2.139	0.178	4.754	0.287	3.487	0.933	7.101
(80, 20)	Firth	-2.424	0.131	6.008	0.140	3.186	0.841	5.726
	Zhang	-2.424	0.132	6.008	0.140	3.183	0.889	6.870

[a] Penalized profile likelihood based interval estimators

Chapter 4

Application

We return to the DES data (Herbst et al., 1971) described in the Introduction. The data are shown in Table 4.1. The study involved 8 cases and 32 controls who were classified according to whether or not they had *in utero* exposure to DES. The study found that 7 of the 8 cases and none of the 32 controls had DES exposure. As shown in Table 4.1, there exists one cell with zero frequency; the data thus present a clear example of quasicomplete separation, with the result that it is impossible to estimate the effect of DES exposure using conventional ML estimation.

Table 4.1: Prenatal exposure to DES among young women with adenocarcinoma of the vagina and among controls

	Case	Control
DES:		
Yes	7	0
No	1	32
Total	8	32

Potential alternatives to ML inference for small samples or data sets with separation include the penalized likelihood approach we studied in this project, and exact logistic regression. Exact logistic regression is based on the conditional distribution of the sufficient statistics for the regression parameters of interest given the observed values for the remaining sufficient statistics (Mehta and Patel, 1995). These conditional distributions are also the basis for exact inference in 2×2 contingency tables (Agresti, 1992). The point estimates can be obtained by maximizing the conditional maximum likelihood (CML). When there is separation in the data and the CML estimate (CMLE) does not exist, Hirji et al. (1989) suggest using the median unbiased estimator (MUE), which is defined as the average of the endpoints of a 50% confidence interval estimator.

To examine the relationship between vaginal cancer and DES exposure, we fit the logistic regression model,

$$\text{logit}(\pi_i) = \text{logit}\{P(y_i = 1|x_i, \beta)\} = \alpha + x_i\beta,$$

where x_i is the DES exposure ($x_i = 1$ for DES exposed and 0 for not exposed).

Here we applied Firth's and Zhang's penalized likelihood approach to the DES data, ignoring matching. For Firth and Zhang logistic regression, we implemented the point estimator, standard errors and 95% penalized profile likelihood confidence interval as described in Section 3.1. The call to the `logistf()` function for Firth logistic regression follows the same structure as the standard functions `lm()` or `glm()`, requiring a data frame and formula for the model specification. The estimates, standard errors and 95% penalized profile likelihood confidence intervals can be extracted using `summary()` function as usual. Below is the sample code for data construction and model fitting.

```
x <- c(rep(0, 33), rep(1, 7))
y <- c(rep(0, 32), rep(1, 8))
DESdata <- as.data.frame(cbind(y,x))

# Firth logistic regression
library(logistf)
DESFirth <- logistf(y ~ x, data = DESdata)
summary(DESFirth)

# Zhang logistic regression
source(ZhangFunctions.R) # see appendix
DESZhang <- logistzCC(formula(y ~ x), DESdata)
summary(DESZhang)
```

For comparison, we also applied exact logistic regression to the data using `PROC LOGISTIC` with the `EXACT` statement in `SAS`. The sample code is as follows:

```
data DESdata;
  input y DES n;
datalines;
1 0 0
1 1 7
0 1 1
0 0 32
;
run;

proc logistic data = DESdata desc;
  freq n;
```

```
  model y = DES;
  exact DES / estimate = both;
run;
```

We also used the `elrm()` function in the `elrm` package (Zamar et al., 2007) to approximate exact logistic regression in R. This is based on MCMC sampling. It requires a collapsed data set with number of successes, covariates combination, and number of trials. For the DES data, the data frame can be constructed as follows:

```
DESdata.elrm <- data.frame(ycount = c(1, 7), DES = c(0, 1), n = c(33, 7))
```

```
  ycount DES n
1      1   0 33
2      7   1  7
```

The first row indicates that there are $n = 33$ subjects who were not exposed to DES (DES = 0) and only one of them (ycount = 1) is case. The second row has similar interpretation.

The logistic model is specified as number of successes out of number of trials, given the covariates. The parameters of interest should also be specified, with all others being considered as nuisance parameters. For DES data, we did 2,200,000 iterations with a 200,000 burnin for a final chain of 2,000,000. The model can be built as follows:

```
library(elrm)
DESexact = elrm(ycount/n ~ DES, interest = ~DES,
               iter = 2200000, burnIn = 200000, data = DESdata.elrm, r = 2)
summary(DESexact)
```

Table 4.2: Estimates, standard errors, p-values and 95% confidence intervals from fitting Firth and Zhang logistic regression, exact logistic regression, and approximate exact logistic regression to data in Table 4.1. The Firth and Zhang p-values are obtained assuming that twice the log-penalized likelihood ratio has a χ^2 distribution with one degree of freedom.

Method	Est.	Std.err.	p-value	95% CI
Firth	5.7838	1.7767	1.4852×10^{-7}	(3.1623, 10.8443)
Zhang	5.7866	1.7605	1.4852×10^{-7}	(3.1657, 10.8472)
exact	4.9364	NA	$<.0001$	$(3.0536, \infty)$
elrm	4.944	NA	0	$(2.6967, \infty)$

The results are presented in Table 4.2. The estimates, standard errors and 95% penalized profile likelihood confidence intervals of the two penalized estimators, Firth and Zhang, are very similar. The estimates based on 2 million MCMC replicates using `elrm()` are close to the estimate obtained from exact logistic regression using SAS. The penalized estimates of about 5.78 are larger than the MUE of 4.94 obtained by the exact methods.

This result is consistent with the simulation results of Heinze and Schemper (2002), which showed that the MUE's from exact logistic regression are pulled towards zero even more than the Firth penalized likelihood estimators for large parameter values (i.e. more data sets with separation). The upper limit of infinity seen in the exact SAS analysis and approximate exact `elrm()` analysis suggests a limitation of this approach relative to the penalized likelihood approaches.

Chapter 5

Concluding Remarks

Many, if not all, epidemiological researchers have encountered separation or "empty cells" in the course of their research when dealing with dichotomous dependent variables in rare diseases. Separation arises most often in situations where small sample sizes and strong relationships are present. Due to the restriction of rare occurrence of disease, small to moderate sample sizes are not uncommon in epidemiological studies. Statistical analysis for small samples is challenging. This is because in small samples, the large-sample approximate distributions used for inference may be unreliable, and the MLEs are substantially biased away from zero. When the small sample problem is coupled with separation, standard ML inference fails as the likelihood does not have a maximum and MLEs do not even exist.

The methods to address this challenge can be classified into two categories, exact logistic regression and penalized logistic regression. For small samples, exact logistic regression is computationally feasible, but still has the limitation that the covariates have to be categorical. The penalized-likelihood method proposed by Firth (1993) was introduced to tackle the separation problem in logistic regression by Heinze and Schemper (2002), as an easy-to-implement solution. The modification of the logistic regression score function to remove the first order bias is equivalent to penalizing the likelihood by Jeffrey's invariant prior. Firth's approach is asymptotically equivalent to standard ML methods in large samples, and superior to them in small samples - the situations in which separation is most likely to be a concern. Penalized logistic regression is therefore an attractive alternative to standard ML approaches when dealing with small to moderate-sized samples, and is preferred over exact logistic regression when there are continuous covariates.

In this project, we have reviewed two penalized likelihood estimators applied in logistic regression: Firth logistic regression developed under a prospective sampling design, and Zhang logistic regression developed under a case-control sampling design. Zhang logistic regression is an extension of Firth's method to case-control data, by introducing a small bias

term into the score function obtained from the semiparametric profile log likelihood, which is derived based on the retrospective case-control sampling design. Based on our simulation study with a single continuous covariate, and data analysis using the DES data, the point and interval estimators from Firth and Zhang logistic regression are virtually identical, for both balanced and unbalanced study designs. Even though there is no formal justification for applying Firth logistic regression to case-control data, it appears to perform as well as Zhang logistic regression which is well justified for case-control data.

For sample sizes of 100 or less, the penalized likelihood methods yield finite parameter estimates that always exist, even in samples in which MLEs do not exist. Thus penalized likelihood methods have advantages over standard ML methods. In most cases, both of the penalized likelihood estimators not only reduce bias but also generally have less variance than the MLEs, and the resulting confidence intervals are generally narrower compared to those of the MLEs. Confidence intervals based on large-sample approximations to Wald statistics can perform badly for both Firth and Zhang logistic regression, especially when the log odds ratio parameter is large. Based on our simulation results, we recommend confidence intervals based on penalized-likelihood-ratio statistics rather than Wald statistics for Firth and Zhang logistic regression.

There are several areas for future work. First, in the project we assume that the data have complete observations; however, missing data are commonly seen in epidemiological research. The application of penalized likelihood estimators to data sets with missing values can be a possible area for future work. Second, although the simulation study has demonstrated that Firth logistic regression performs as well as Zhang logistic regression in case-control data, it would be nice if the application of Firth logistic regression to case-control data could be theoretically justified. Third, Zhang logistic regression is theoretically justified for case-control data, but has not been programmed in major software packages, and so is inconvenient to apply. When implementing the method, we found it is hard to include more than one covariate, because of numerical instabilities in calculating the penalized profile likelihood. Specifically, for a covariate effect of interest, evaluating the penalized profile likelihood at a given point involves maximizing a penalized likelihood over the nuisance parameters. We found that this intermediate maximization step could fail to converge when the nuisance parameters included the effects of other covariates besides an intercept term. This is also one of the reasons that why we use a single covariate in our simulation study.

Bibliography

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153.

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

Bull, S., Mak, C., and Greenwood, C. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis*, 39(1):57–74.

dos Santos Silva, I. (1999). *Cancer Epidemiology: Principles and Methods*. IARC Press, Lyon, France.

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

Heinze, G., Ploner, M., Dunkler, D., and Southworth, H. (2013). *logistf: Firth's bias reduced logistic regression.*

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

Herbst, A. L., Ulfelder, H., and Poskanzer, D. C. (1971). Adenocarcinoma of the vagina. association of maternal stilbestrol therapy with tumor appearance in young women. *The New England Journal of Medicine*, 284(15):878–881.

Hirji, K. F., Tsiatis, A. A., and Mehta, C. R. (1989). Median unbiased estimation for binary data. *The American Statistician*, 43(1):7–11.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*. Wiley series in probability and statistics. John Wiley & Sons, Inc. A Wiley-Interscience Publication, New York, Chichester, Weinheim.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 186(1007):453–461.

Ma, C., Blackwell, T., Boehnke, M., and Scott, L. J. (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology*, 37(6):539–550.

Mehta, C. R. and Patel, N. R. (1995). Exact logistic regression: theory and examples. *Statistics in Medicine*, 14(19):2143–2160.

Nemes, S., Miao, J. J., Genell, A., and Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, 9(1):1–5.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and control studies. *Biometrika*, 66(3):403–411.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325.

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618.

Zamar, D., McNeney, B., and Graham, J. (2007). elrm: Software implementing exact-like inference for logistic regression models. *Journal of Statistical Software*, 21(3):1–17.

Zhang, B. (2006). Bias-corrected maximum semiparametric likelihood estimation under logistic regression models based on case-control data. *Journal of Statistical Planning and Inference*, 136(1):108–124.

Appendix A

Code

```
# We call our implementation of Zhang's modified score approach logistzCC,
    because it is like Heinze's logistf (Firth logistic regression), but
    specifically for case-control data.
## ------------------------------------------------------------------------
logistzCC = function(formula, data, init = NULL,
                     conf.level = 0.95, verbose.output = FALSE) {
  # Input:
  # - formula is the model formula
  # - data is a data frame that includes the variables named in formula
  # - conf.level is the level for confidence intervals
  # - init (optional) is the initial regression parameter values to use
  # - verbose.output: Should we return the output of the root-finding
  #   algorithm?

  # 1. Set up the data that modscore will need.
  mf = model.frame(formula,data)
  D = model.response(mf) # response variable
  X = model.matrix(formula,data) # includes a column for the intercept
  n=length(D); n1 = sum(D==1); n0 = sum(D==0); rho = n1/n0

  # 2. Define modscore() within logistzCC().
  modscore = function(theta) {
    pi = expit(log(rho) + as.numeric(X%*%theta))
    uu = U(pi,D,X)
    mu.t = mu.tilde(pi,X)
    D.t = D.tilde(pi,X,n1)
    mod  = 0.5*(mu.t + (1-rho)*D.t) # Zhangs's first modified score function U_{M1}
    return(uu-mod)
  }

  # 3. Call the root finder nleqslv() on modscore(), with initial value init if
      passed by the user, or a vector of zeros if not.
  require(nleqslv)
  if(is.null(init)) {init = rep(0,ncol(X))}
  root.info = nleqslv(init,modscore)
  estimate = root.info$x; names(estimate) = colnames(X)
```

```
  # 4. Covariance, SEs and Wald CI
  pi.final = expit(log(rho) + as.numeric(X%*%estimate))
  var = logistVar(pi.final, X)
  ses = sqrt(diag(var))
  Wald.CI = Waldci(estimate, ses, conf.level)

  # 5. PPL CI
  if(ncol(X)==2) {
    PPL.CI = PPLci(logPPL.ZhangH,estimate[2],X,D,conf.level)
  } else {
    stop("Model has ",ncol(X)-1," covariates. Only 1 allowed for PPL CIs.")
  }
  colnames(PPL.CI) = colnames(Wald.CI)
  rownames(PPL.CI) = rownames(Wald.CI)
  # Return the results.
  out=list(coefficients=estimate,Wald.CI=Wald.CI,PPL.CI = PPL.CI,var=var)
  if(verbose.output) out$root.info = root.info
  return(out)
}
## ------------------------------------------------------------------------
# Functions needed by logisfCC
## ------------------------------------------------------------------------
expit = function(t) { return(exp(t)/(1+exp(t)))}
## ------------------------------------------------------------------------
U = function(pi,d,X) {
  return(colSums((d-pi)*X))
}
## ------------------------------------------------------------------------
D.tilde = function(pi,X,n1) {
  return(colSums(pi*(1-pi)*X)/n1)
}
## ------------------------------------------------------------------------
mu.tilde = function(pi,X) {
  pp1 = ncol(X) # p+1
  n = nrow(X)
  s.inv = solve(S.tilde(pi,X))
  out = rep(0,pp1)
  for(m in 1:pp1) {
    out[m] = tr(Lambda.tilde(pi,X,m)%*%s.inv)
  }
  return(out)
}
tr = function(mat) { return(sum(diag(mat)))}
## ------------------------------------------------------------------------
Lambda.tilde = function(pi,X,m){
  pp1 = ncol(X) # p+1
  n = nrow(X)
  out = matrix(0,nrow=pp1,ncol=pp1)
  for(i in 1:n) {
    H.im = X[i,]%*%t(X[i,])*X[i,m]
    out = out - pi[i]*(1-2*pi[i])*(1-pi[i])*H.im
  }
  return(out/n)
```

```
}
## ------------------------------------------------------------------------
S.tilde = function(pi,X) {
  pp1 = ncol(X) # p+1
  n = nrow(X)
  out = matrix(0,nrow=pp1,ncol=pp1)
  for(i in 1:n) {
    H.i1 = X[i,]%*%t(X[i,])
    out = out + pi[i]*(1-pi[i])*H.i1
  }
  return(out/n)
}

logistVar= function(pi.final,X) {
  n = nrow(X)
  return(solve(n* S.tilde(pi.final,X)))
}

Waldci = function(estimate, ses, conf.level) {
  crit.val = qnorm((1-conf.level)/2, lower.tail = FALSE)
  Wald.CI = cbind(estimate-1.96*ses, estimate+1.96*ses)
  colnames(Wald.CI) = c("CI.Lower", "CI.Upper")
  rownames(Wald.CI) = names(estimate)
  Wald.CI = Wald.CI[-1,,drop=FALSE] # CI for intercept not meaningful (?)
  return(Wald.CI)
}
## ------------------------------------------------------------------------
# PPL confidence intervals:

PPLci = function(logPPLfunc, betahat, X, D, conf.level, alpha.lim=c(-100, 100)){
  # Inputs:
  # - logPPLfunc is the logPPL function to use (currently only logPPL.ZhangH)
  # - betahat is the maximum PPL estimate of the log OR beta
  # - X is the design matrix
  # - D is the vector of disease status
  # - conf.level is the confidence level for the interval estimator

  # Output:
  # - the PPL confidence interval

  beta.lower = c(-50, betahat); beta.upper = c(betahat, 50)
  betaProf.max = betaProf(betahat, X, D, alpha.lim, logPPLfunc)
  crit.val = qchisq(conf.level, df=1)
  objfunc = function(beta) {
    bpdiff = betaProf(beta, X, D, alpha.lim, logPPLfunc) - (betaProf.max -
        crit.val/2)
    return(bpdiff^2)
  }
  ci.lower = optimize(f = objfunc, interval = beta.lower)$minimum
  ci.upper = optimize(f = objfunc, interval = beta.upper)$minimum
  # format CI as a matrix to match formatting of Wald CI
  ci = matrix(c(ci.lower, ci.upper), nrow = 1, ncol = 2)
  return(ci)
```

```
}

betaProf = function(beta, X, D, alpha.lim = c(-100,100), logPPLfunc) {
  # for fixed beta maximize logPPLfunc over alphas in alpha.lim
  oo = optimize(f = logPPLfunc, interval = alpha.lim, beta = beta, X = X, D = D,
                maximum=TRUE)
  return(oo$objective)
}

logPPL.ZhangH = function(alpha, beta, X, D) {
  coef = c(alpha,beta)
  p = expit(as.numeric(X%*%coef))
  n1 = sum(D)
  return(logPL(coef, X, D) + log.JeffH(p, X) - sum(p)/(2*n1))
}
# The log profile likelihood is shown by Zhang to have the same form as the log
    likelihood for logistic regression based on prospective data.
logPL = function(coef, X, D) {
  return(sum(D*(as.numeric(X%*%coef)) - log(1 + exp(as.numeric(X%*%coef)))))
}
log.JeffH = function(p,X) {
  # Can show that the Hessian is X^T W X where W is a diagonal matrix of variance
      terms p*(1-p).
  vv = p*(1-p)
  Hess = t(X)%*% diag(vv) %*% X
  return(0.5*log(det(-1*Hess)))
}

# Data simulator
simdat = function(n0, n1, beta) {
  # Dist'n of covariate x in controls is standard normal.
  # Can show it is normal with mean beta and sd 1 in cases.
  case = c(rep(1, n1), rep(0, n0))
  x = c(rnorm(n1, mean = beta, sd = 1), rnorm(n0, mean=0, sd=1))
  return(data.frame(case = case, x = x))
}
```